FieldStone

C. Thieulot

If you find anything in this document useful for your research please cite it as follows (*please include the doi!*):

```
@article{fieldstone,
author = "Cedric Thieulot",
title = "{Fieldstone: a computational geodynamics (self-)teaching tool}",
year = "2023",
doi = "10.5194/egusphere-egu23-14212"}
```

Why do I have to promise where I am going while I am not there yet?

You can't google something you don't know exists.

You can be correct or you can get stuff done.

Contents

Chapter 1

Introduction

1.1   Philosophy

This document was written with my students in mind, i.e. 3rd and 4th year Geology/Geophysics students at Utrecht University. I have chosen to use as little jargon as possible unless it is a term that is commonly found in the geodynamics literature (methods paper as well as application papers). There is no mathematical proof of any theorem that may be mentioned but I will try to refer to the appropriate sources, i.e. generic Numerical Analysic, Finite Element and Linear Algebra books. If you find that this books lacks references to Sobolev spaces, Hilbert spaces, and other spaces, this book is just not for you.

The codes I provide here are by no means optimised as I have chosen code readability over code efficiency. I have also chosen to avoid resorting to multiple code files or even functions in order to favour a sequential reading of the codes. These codes are not designed to form the basis of a real life application: Existing open source highly optimised codes shoud be preferred, such as ASPECT [776, 593], CITCOMS [1523, 1521], LAMEM [724], PTATIN [897, 894], PYLITH [1], ... (see Appendix **??**).

Concerning figures I have consciously decided not to place them inside *figure* LaTeXenvironments since it does not allow for complete control over where they end up. Instead they are inserted when they are needed in the text.

All kinds of feedback is welcome on the text (grammar, typos, ...), on the text, the equations or on the code(s). You will have my eternal gratitude if you wish to contribute an example, a benchmark, a cookbook.

All the python scripts and tex files are freely available at

$$https://github.com/cedrict/fieldstone$$

This document is available at:

$$https://cedrict.github.io/$$

Disclaimer: there are many things in this huge document I probably do not fully understand, or that I am simply wrong about. I sometimes write open questions in the text about such things. My commitment is to revisit this document time and time again, until it is 99% correct. This is not a book, it has not been edited by anybody. It is not perfect in any way. I nevertheless hope it will be useful to many in the long run.

1.2 ambition & motivation

I wish to provide the community with:

- a ginormous bibliography data base - simply search the pdf for keywords. The LaTeX bib file[1] is also available next to the manual.tex file on github;

- a go-to document for anybody who wants to know more about a particular topic in computational geodynamics;

- a useful teaching tool for researchers, teachers, students and PhD students alike;

- small, readable, educative codes.

1.3 Acknowledgements

I have benefitted from many discussions, lectures, tutorials, coffee machine discussions, debugging sessions, conference poster sessions, etc ... over the years. I wish to name these instrumental people in particular and in alphabetic order: Wolfgang Bangerth, Jean Braun, Taco Broerse, Rens Elbertsen, Philippe Fullsack, Menno Fraters, Anne Glerum, Timo Heister, Dave May, Robert Myhill, John Naliboff, E. Gerry Puckett, Melchior Schuh-Senlis, Michael Tetley, Lukas van de Wiel, Arie van den Berg, Eric van den Hoogen, Tom Weir, and the whole ASPECT family/team.

1.4 About the author

I have BSc in mathematics, and an MSc diploma in physics (with a specialization in musical acoustics [349]). I did my PhD at the university of Groningen (The Netherlands) titled *Thermodynamically consistent fluid particle modelling of phase separating mixtures*[2]. Although half of the thesis deals with the re-derivation of the Navier-Stokes equations for such systems[406], the second half is concerned with the implementation of these equations with the Smoothed Particle Hydrodynamics method [1350, 1351, 1349].

I then taught physics and programming at the University of Rennes (France) for a year, after which I did a 2-year post-doc with Prof. J. Braun[3] in the Geosciences department. I then did a 4-year post-doc with prof. R. Huismans[4] at the University of Bergen (Norway), followed by a 3-year post-doc with profs. T. Torsvik and W. Spakman at the Utrecht University (The Netherlands). Since June 2015 I am assistant professor there in the 'Mantle dynamics & theoretical geophysics' group.

1.5 Installing packages

---

[1] https://github.com/cedrict/fieldstone/blob/master/biblio_geosciences.bib
[2] http://cedricthieulot.net/thesis.html
[3] https://www.gfz-potsdam.de/en/staff/jean-braun/
[4] https://folk.uib.no/huismans/

Python

If numpy, scipy or matplotlib are not installed on your machine, here is how you can install them:

```
sudo apt install python3-numpy
sudo apt install python3-scipy
```

To install the umfpack solver (check?):

```
pip install --upgrade scikit-umfpack --user
```

If you need to install pip:

```
sudo apt install python3-pip
```

Julia

In order to have vim supporting the Julia language, do

```
git clone git@github.com:JuliaEditorSupport/julia-vim.git
```

and copy the content of the julia-vim folder in the .vim folder. That's it.

LaTeX

To install siunitx package:

```
sudo apt -y install texlive-science
```

To install additional fonts:

```
sudo apt-get install texlive-fonts-extra
```

To install biber package:

```
sudo apt install biber
```

1.6   What is a (real) fieldstone?

whatisafieldstone.tex

Taken from https://en.wikipedia.org/wiki/Fieldstone

Simply put, it is a stone collected from the surface of fields where it occurs naturally. It also stands for the bad acronym: *fi*nite *el*ement *d*eformation of *stone*s which echoes the primary application of these codes: geodynamic modelling.

1.7 Why the Finite Element method?

The Finite Element Method (FEM) is by no means the only method to solve PDEs in geodynamics, nor it is necessarily always the best one. Other methods are employed very successfully, such as the Finite Difference Method (FDM), the Finite Volume Method (FVM), and to a lesser extent the Discrete Element Method (DEM) [1320, 383, 384, 456, 685], the Lattice-Boltzmann method [634], the Rigid Element Method [797], or the Element Free Galerkin Method (EFGM) [564]. I have been using FEM since 2008 and I do not have real experience to speak of in FVM or FDM (except for chapter 11) so I concentrate in this book on what I know best.
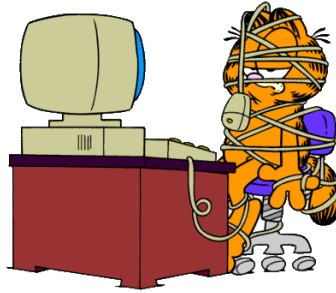
1.8 Notations

Scalars such as temperature, density, pressure, etc ... are simply obtained in LaTeX by using the math mode, e.g. T, ρ, p. Although it is common to lump vectors and matrices/tensors together by using bold fonts, I have decided in the interest of clarity to distinguish between those: vectors are denoted by an arrow atop the quantity, e.g. $\vec{v}$, $\vec{g}$, while matrices and tensors are in bold $\boldsymbol{M}$, $\boldsymbol{\sigma}$, etc ...
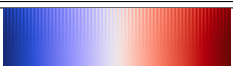
Also I use the $\cdot$ notation between two vectors to denote a dot product $\vec{u} \cdot \vec{v} = u_i v_i$ or a matrix-vector multiplication $\boldsymbol{M} \cdot \vec{a} = M_{ij} a_j$. If there is no $\cdot$ between vectors, it means that the result $\vec{a}\vec{b} = a_i b_j$ is a matrix (it is a dyadic product[5]). Case in point, $\vec{\nabla} \cdot \vec{v}$ is the velocity divergence while $\vec{\nabla}\vec{v}$ is the velocity gradient tensor.

1.9 Colour maps for visualisation

In an attempt to homogenise the figures obtained with ParaView, I have decided to use a fixed colour scale for each field throughout this document. These colour scales were obtained from `https://peterkovesi.com/projects/colourmaps` and are Perceptually Uniform Colour Maps [771].

---

[5]`https://en.wikipedia.org/wiki/Dyadics`

Field	colour code	
Velocity/displacement	CET-D01A	
Pressure	CET-L17	
Velocity divergence	CET-L01	
Density	CET-D03	
Strain rate	CET-R2	
Viscosity	CET-R3	
Temperature	CET-D09	
stress	CET-L18	
Spin tensor	CET-R1	
Composition field	CET-CBD1	
Gravity acceleration	vik	
Gravity potential	roma	
Vorticity	CET-L12	
Stream function	CET-D02	

vik and roma are available at `http://www.fabiocrameri.ch/colourmaps.php`. See also Crameri *et al.* (2020) for a discussion about the misuse of colour is science communication.

1.10 How my bibliography works

There is a single (large) bibliography file for this document:

`biblio_geosciences.bib`

If the paper is a single-author paper, say by Garfield[6], published in 1978[7], its code in my bibliography file is *garf78* (i.e. the first four letters of the name, followed by the two digits of the publication year).

If the paper was written by two authors, say Garfield and Odie, in 1987, its code will be *gaod87*, i.e. the first two letters of the first author followed by the two first letters of the second author followed by two digits.

If the paper was written by three or more authors, say Garfield, Odie, John and Irene in 2003, its code will be *gaoj03*, i.e. the first two letters of the first author followed by the first letter of the second author, the first letter of the third author and the year.

If multiple papers are published the same year by the same authors, I simply append a,b,c... to the above rules.

---

[6]This is just an example

[7]May be not, after all, since Garfield the cat was born in 1978

Remark. *Dutch names such as 'van Hunen' or 'van den Berg' are classified under letter 'v', not 'h' or 'd' nor 'b'.*

In Section **??** I present an analysis of the literature in my bib file broken down per year and/or journal.

1.11 How to download a single stone

Say you wish to only download a single python program, for example stone 3. You then go to `https://github.com/cedrict/fieldstone` and click on `python_codes`, then on `fieldstone_03` and then on `stone.py`.

Then click on Raw:
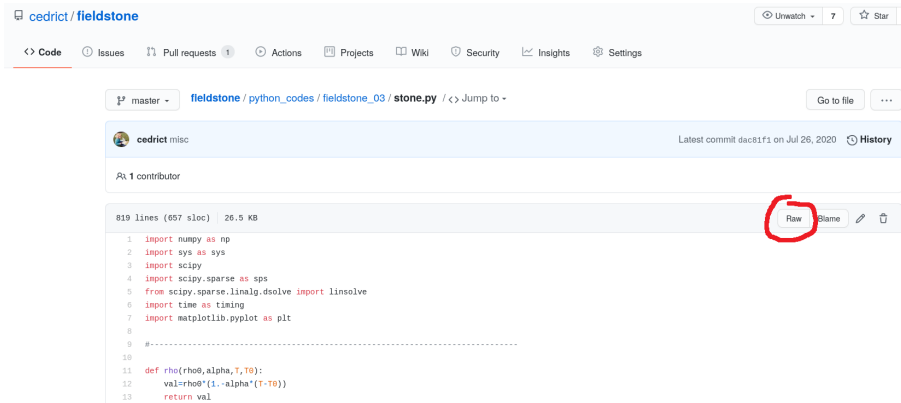
And then copy the address in the address bar:
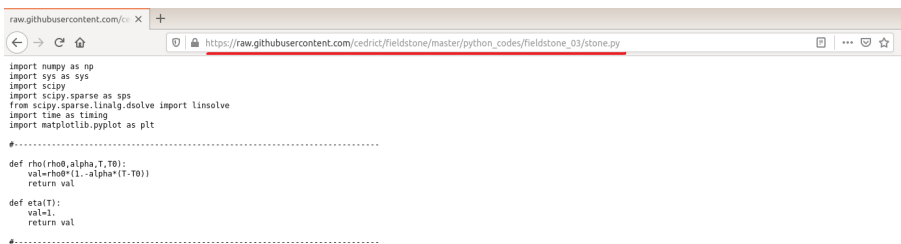
Finally, in the terminal of your Linux/Apple computer type

`wget https://raw.githubusercontent.com/cedrict/fieldstone/master/python_codes/fieldstone_0`

1.12 Oldies but goodies

I hereunder show a few figures taken from early-ish geodynamics FEM papers.
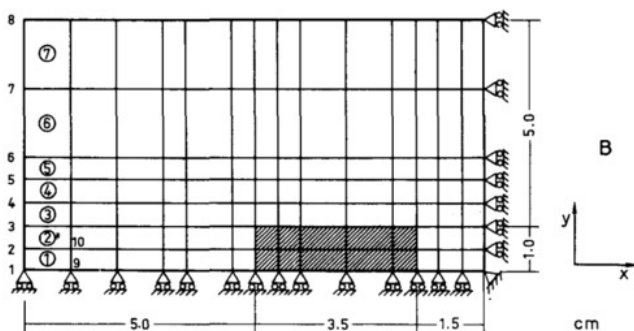
Model a boudinage structure - Stephansson and Berner [1292] (1971)
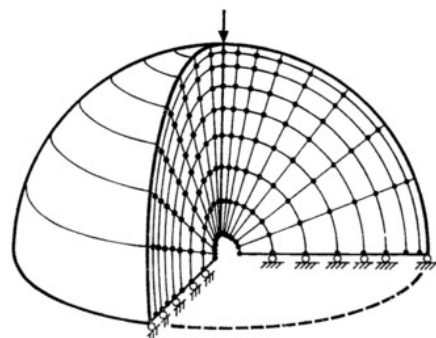
Crustal Structure from Surface Load Tilts - Beaumont and Lambert

[67] (1972)
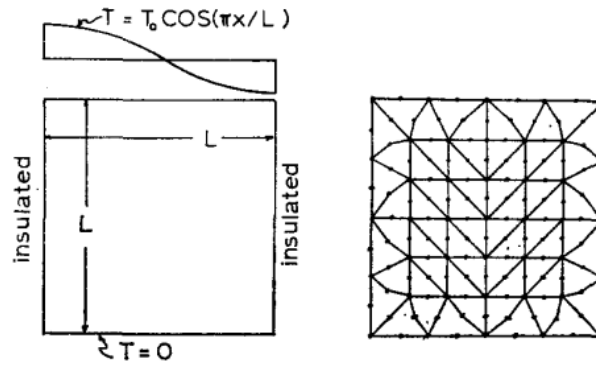
Mantle convection in a square domain, thermo-mechanically coupled! Sato and Thompson [1190] (1976)
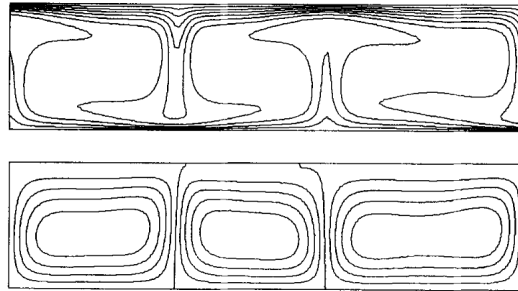
Mantle convection in a rectangular domain - Lux, Davies, and Thomas [863] (1979)
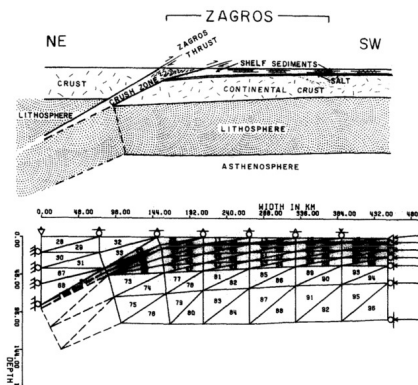
Finite element modelling of lithosphere deformation: the Zagros collision orogeny - Bird [98] (1978)
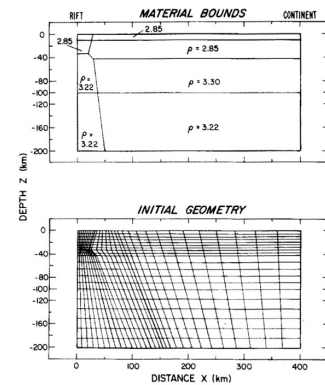
Thermal regimes, mantle diapirs and crustal stresses of continental rifts - Bridwell and Potzick [165] (1981)
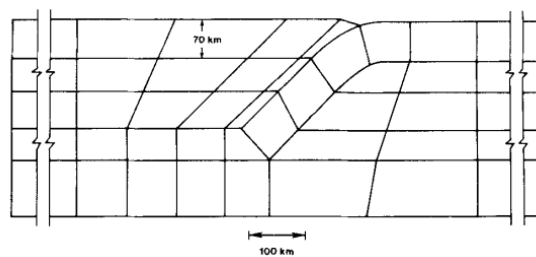
Numerical models of subduction and forearc deformation - Tharp [1338] (1985)
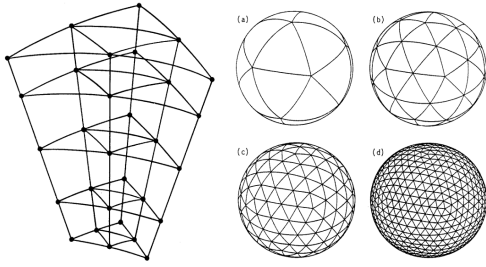
Three-Dimensional Treatment of Convective Flow in the Earth's Mantle -
Baumgardner [62] (1985)

CRUSTAL DEFORMATION FIELD

a) short λ

b) long λ

Moho

c) combined

Moho

Moho

Lithospheric necking: a dynamic model for rift morphology - Zuber,
Parmentier, and Fletcher [1550] (1986)

FINITE ELEMENT GRID FOR MODELS 1 TO 6

Distance (km)
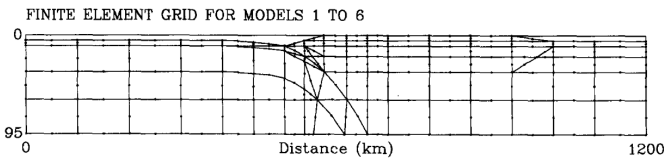
Plate boundary forces at subduction zones and trench-arc compression -
Bott, Waghorn, and Whittaker [124] (1989)

a)Initial grid   t=0 m.y.

A   B $T=0$ C   D

$A \neq 0$ Quartz ρ_c crust
$A = 0$ Feldspar

Moho   Olivine   No tilt

$\frac{dT}{dz} = 0$ ρ_m lithosphere σ_{xx}

$k\frac{dT}{dz} = Q^*$ ρ_a asthenosphere

Distance (km)

Relation between flank uplifts and the breakup unconformity at rifted
continental margins - Braun and Beaumont [149] (1989)
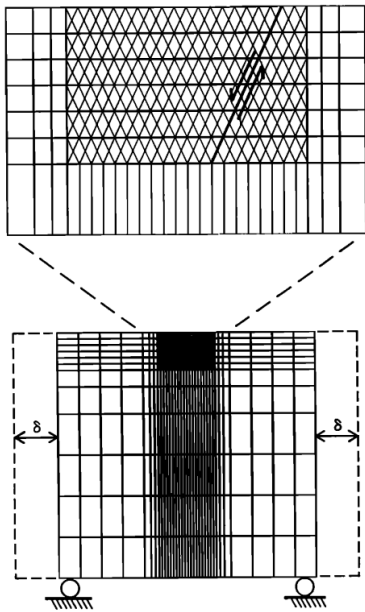
Mechanics of graben formation in crustal rocks -
Melosh and Williams Jr [912] (1989)

FINITE ELEMENT GRID FOR MODELS 3 AND 4

Distance (km)

FINITE ELEMENT GRID FOR MODELS 7 AND 8

Distance (km)
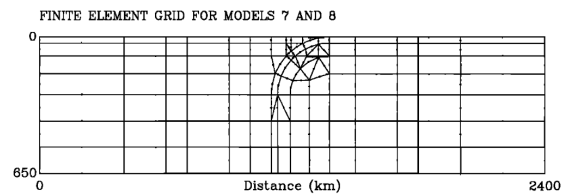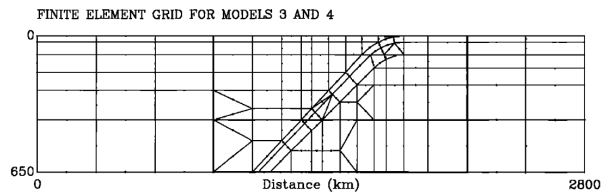
Stresses and plate boundary forces associated with subduction plate margins - [1453]
(1992)
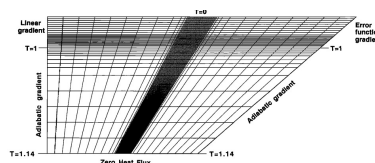
Temperature field in subduction zones - Davies and Stevenson [336] (1992)
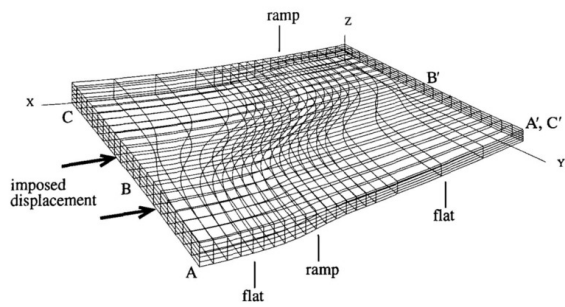
3D numerical modeling of compressional orogenies: Thrust geometry
and oblique convergence - Braun [147] (1993)
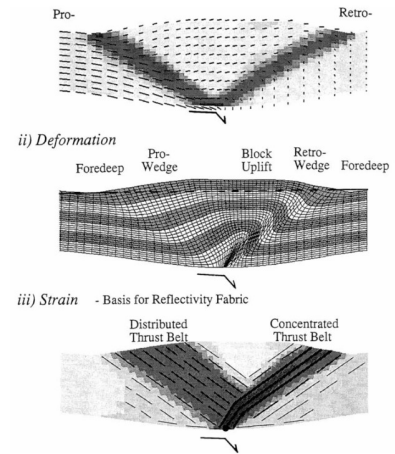
Crustal-scale compressional orogens - Beaumont and Quinlan [68]
(1994)
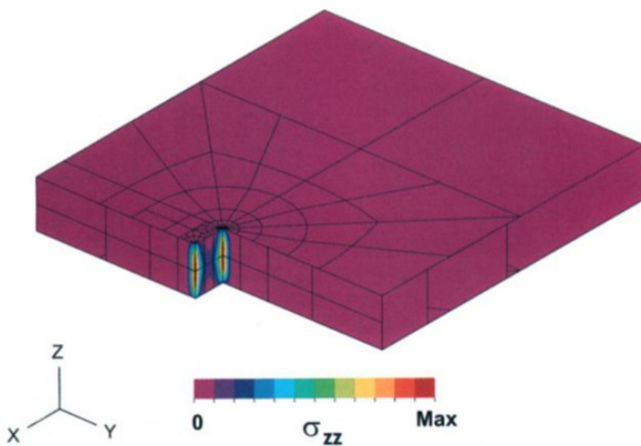
Modeling of pull-apart basins - Katzman, Brink, and Lin [718] (1995)
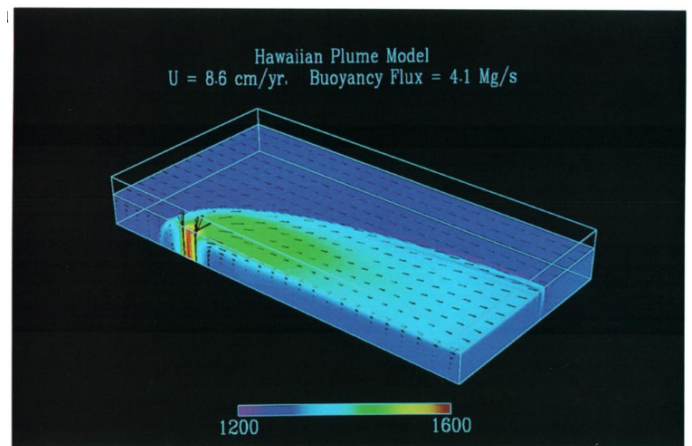
Plume-lithosphere interaction - Ribe and Christensen [1134] (1994)
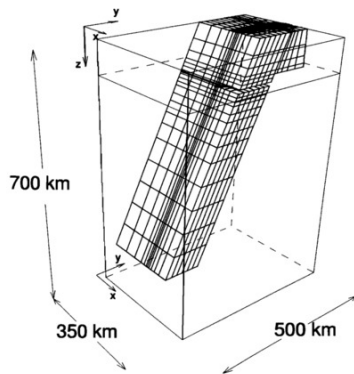
3D numerical modeling of detachment of subducted lithosphere -
Yoshioka and Wortel [1491] (1995)
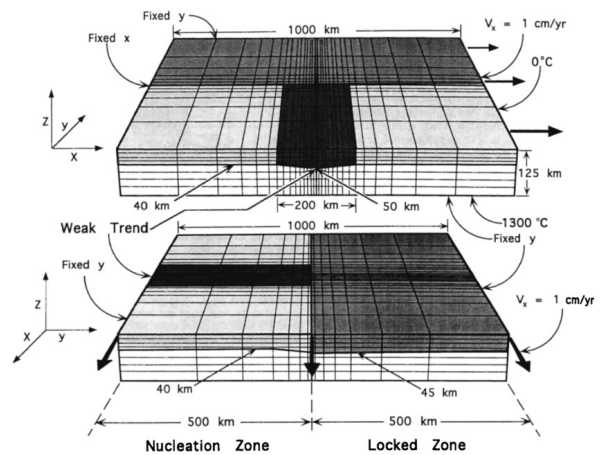
3D dynamical model of continental rift propagation and margin
plateau formation - Dunbar and Sawyer [374] (1996)
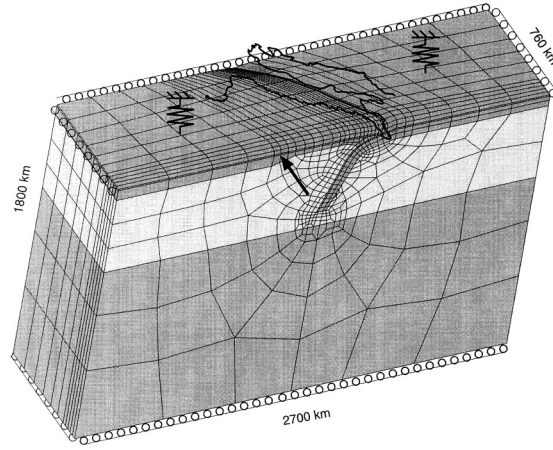
Model geometry, boundary conditions and 3-D finite element mesh used in the calculations. The circles denote a free-slip condition. The arrow denotes the velocity applied in some calculations to the southern boundary of the Tyrrhenian domain to simulate the motion of the African plate. The springs represent the buoyant restoring force applied at the surface - Negredo, Sabadini, Bianco, and Fernandez [989] (1999)

📇 Relevant Literature: Gartling [463] (1978), Anderson and Bridwell [21] (1980), Melosh and Raefsky [911] (1980), Bridwell and Anderson [164] (1980), England [397] (1982), Tharp [1338] (1985), Schubert and Anderson [1218] (1985), England and Houseman [398] (1986), Moretti and Froidevaux [958] (1986), Zuber and Parmentier [1549] (1986), Bott, Waghorn, and Whittaker [124] (1989).

Chapter 2

Resources and more

2.1   Videos

- `https://youtu.be/aLJMDn_2-d8` [10min]
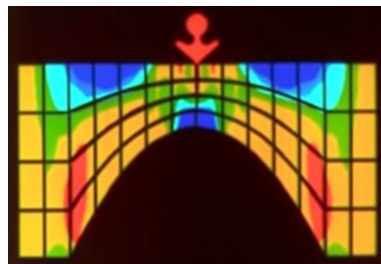
- `https://youtu.be/j2_dJY_mIys` [10min] Smarter Every Day channel

- `https://youtu.be/X4zd4Qpsbs8` [2min] Reversible Stokes flow (cylinder + dye)

- `https://youtu.be/wzcVT0oZJkg` [12min] Boring Through The Earth's Crust

- `https://youtu.be/GHjopp47vvQ` [18min] Understanding the Finite Element Method
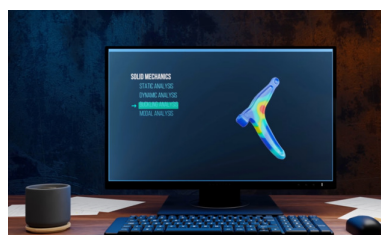
- `https://www.esa.int/Applications/Observing_the_Earth/GOCE/Gravity_mission_still_unearthing_hidden_secrets` [3min] GOCE helps create new model of crust and upper mantle

- `https://youtu.be/_5q8hzF9VVE` [12min] Continental drift (Wegener theory)

- `https://youtu.be/ZTRu620bIsE` [12min] Plate tectonics

- `https://youtu.be/V_zsD8vXyik` [5min] Heat tranfer

- `https://youtu.be/q6S03qA0-n4` [4min] What is sea level? (geoid)

- `https://www.youtube.com/watch?v=FTh694ugfFU` [2min] Bart Root - Huygens pendulum (Gouden Kalf) to measure gravity

- `https://youtu.be/kijOMOGDfRo?si=qmM8lnyJ5XTFw83n` [35min] Rob Moucha - Coupling Geodynamic and Landscape Evolution Models

- `https://youtu.be/eNhJY-R3Gwg?si=Fbs_4kK5yqAJMi2b` [6min] minutephysics - General Relativity Explained in 7 Levels of Difficulty

2.1.1   Dave Whipp videos

They are available at `https://www.youtube.com/@helsinkiuniversitygeodynam6511/videos`. Click on the film icon to launch video on YouTube.

- set 2: Kinematics of plate tectonics

  1. Divergent plate boundaries [12min] 🎞
  2. Convergent plate boundaries [9min] 🎞
  3. Transforms and the Wilson cycle [7min] 🎞
  4. Plate motions on a flat Earth [10min] 🎞
  5. Triple junctions [9min] 🎞
  6. Plate motions on a sphere [9min] 🎞

- set 3: forces and stresses

  1. Forces [9min] 🎞
  2. Surface stresses [14min] 🎞
  3. Stresses in 2D [11min] 🎞
  4. Stresses in 3D [11min] 🎞

- set 4: measuring stress and strain

  1. Global and regional stress measurements [9min] 🎞
  2. How is stress measured? [11min] 🎞
  3. What is strain? [9min] 🎞
  4. Normal and shear strains [9min] 🎞
  5. Historical strain measurements [11min] 🎞
  6. Modern strain measurements [10min] 🎞

- set 5: basics of elasticity

  1. The elastic Earth [8min] ⬚⬚⬚⬚
  2. Elasticity in terms of stress [8min] ⬚⬚⬚⬚
  3. Uniaxial stress [7min] ⬚⬚⬚⬚
  4. Uniaxial strain [7min] ⬚⬚⬚⬚
  5. Pure and simple shear [5min] ⬚⬚⬚⬚
  6. Isotropic stress [7min] ⬚⬚⬚⬚

- set 6: flexure of the lithosphere

  1. Flexure of an elastic plate [6min] ⬚⬚⬚⬚
  2. Balancing forces and torques [18min] ⬚⬚⬚⬚
  3. Deflection of a simple elastic plate [8min] ⬚⬚⬚⬚
  4. Flexure of the lithosphere [11min] ⬚⬚⬚⬚
  5. Examples of flexure of lithospheric plates [12min] ⬚⬚⬚⬚

- set 7: heat conduction and production

  1. The importance of heat [7min] ⬚⬚⬚⬚
  2. Heat conduction [7min] ⬚⬚⬚⬚
  3. Heat production [9min] ⬚⬚⬚⬚
  4. Heat transfer in 1d [10min] ⬚⬚⬚⬚
  5. Distribution of heat producing elements [6min] ⬚⬚⬚⬚
  6. Influence of topography [12min] ⬚⬚⬚⬚

- set 8: thermal processes in the lithosphere

  1. Time-dependent heat transfer [5min] ⬚⬚⬚⬚
  2. Examples of time dependence [9min] ⬚⬚⬚⬚
  3. Heat transfer by advection [12min] ⬚⬚⬚⬚
  4. Examples of advection [12min] ⬚⬚⬚⬚
  5. The Peclet number [6min] ⬚⬚⬚⬚
  6. Thermal structure of active tectonic areas [10min] ⬚⬚⬚⬚

- set 9: basic of fluid mechanics

  1. Introduction to fluid mechanics [8min] ⬚⬚⬚⬚
  2. 1D channel flows I [9min] ⬚⬚⬚⬚
  3. Channel flows in 1D II [9min] ⬚⬚⬚⬚
  4. Channel flow in salt tectonics [6min] ⬚⬚⬚⬚
  5. Asthenospheric counterflow [5min] ⬚⬚⬚⬚
  6. Pipe flow [8min] ⬚⬚⬚⬚

- set 10: plate-driving forces

min ▭▭▭

- set 11: brittle deformation and faulting

  min ▭▭▭

- set 12: viscous deformation, strength of the lithosphere

  min ▭▭▭

- set 13: climatic, geomorphic and geodynamic processes

  min ▭▭▭

2.1.2 Clint Conrad videos

- Plate Tectonics: Linking Surface Geology to Earth's Deep Interior [35min] ▭▭▭

- Models of mantle convection [1h30] ▭▭▭

- Mantle flow for the present day [1h30] ▭▭▭

- Mantle flow for Earth history [54min] ▭▭▭

- 50 years of plate tectonics. But what is the driving force? [24min] ▭▭▭

2.1.3 Wolfgang Bangerth videos

I need to compile a list

2.1.4 Fluid Mechanics 101 channel videos

- SIMPLE Algorithm Course. [21min] ▭▭▭

- Conjugate Gradient for CFD (Part 1): Background and Steepest Descent [45min] ▭▭▭

- Conjugate Gradient for CFD (Part 2): Optimum Distance and Directions [34min] ▭▭▭

- Conservative, Advective & Material Derivative forms of the N-S Equations [32min] ▭▭▭

- Gauss-Seidel Method in CFD [29min] ▭▭▭

- The Courant (CFL) Number [28min] ▭▭▭

- Lagrangian Particle Tracking [29min] ▭▭▭

- The Boussinesq Approximation for Bouyancy Driven (Natural Convection) Flow [18min] ▭▭▭

- The Energy Equation for Solids and Fluids in CFD [31min] ▭▭▭

- The Finite Volume Method in CFD [24min] ▭▭▭

- Eulerian Multi-Phase Modelling [25min] ▭▭▭

2.2 Online resources

`https://software-carpentry.org/lessons/` core lessons:

- The Unix Shell

- Version Control with Git

- Programming with Python

- Plotting and Programming in Python

- Programming with R

- R for Reproducible Scientific Analysis

2.3 Textbooks

[483]	[484]	[659]	[363]
[79]	[80]	[1216]	[960]
[1250]	[789]	[992]	

- *Numerical modeling of Earth Systems* by Thorsten W. Becker and Boris J. P. Kaus, `http://www-udc.ig.utexas.edu/external/becker/teaching-557.html`

- *Myths & Methods in Modeling* by M. Spiegelman,
  https://www.ldeo.columbia.edu/~mspieg/mmm/

- *Computational Science I* by Matthew G. Knepley,
  https://cse.buffalo.edu/~knepley/classes/caam519/Syllabus.html

- *Introduction to Numerical Methods for Variational Problems* by Hans Petter Langtangen and Kent-Andre Mardal,
  https://hplgit.github.io/fem-book/doc/pub/book/pdf/fem-book-4print.pdf

2.4 Learning python

- Introduction to Python & Data (Utrecht University)
  https://utrechtuniversity.github.io/workshop-introduction-to-python/

- This is a joint course for geography and geology students at the Department of Geosciences and Geography, University of Helsinki
  https://geo-python.github.io/site/

- Python for Earth Science Students (University of California San Diego)
  https://github.com/ltauxe/Python-for-Earth-Science-Students

2.5 Published peer-reviewed literature

In the statistics_biblio folder there is a bash script which analyses the .bib file of FieldStone. I also have another script on my computer which analyses the pdf files in my literature folder:
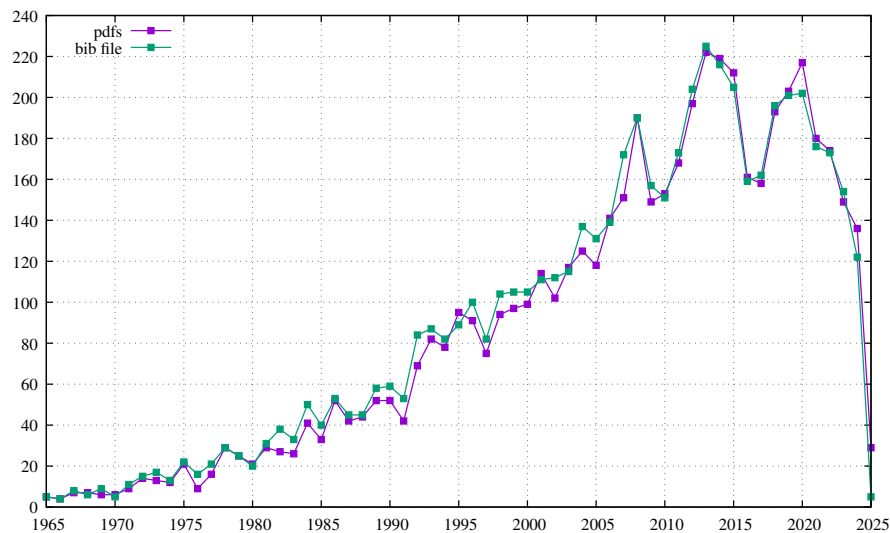
Evolution of the number of references cited in FieldStone per year. The purple line tracks the pdf files in my folder while the green line accounts for the entries in my .bib file.

In what follows I have selected a subset of all journals present in my bib file and report on the number of articles per year for each of them:
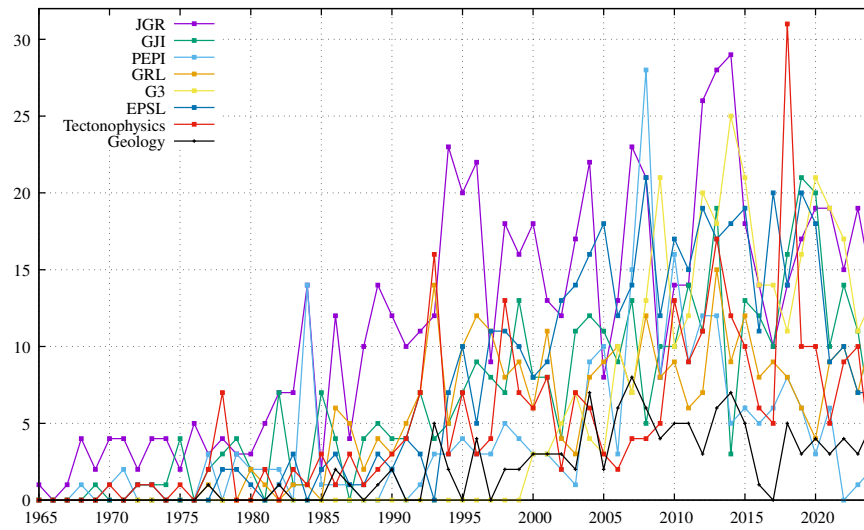
Evolution of the number of references cited in FieldStone per year per journal.

Unfortunately this figure is not as readable as I had expected so I further break it down per journal:

- **Solid Earth**[1] is a not-for-profit journal that publishes multidisciplinary research on the composition, structure, and dynamics of the Earth from the surface to the deep interior at all spatial and temporal scales. It is published by EGU/Copernicus Publications and started in 2010.
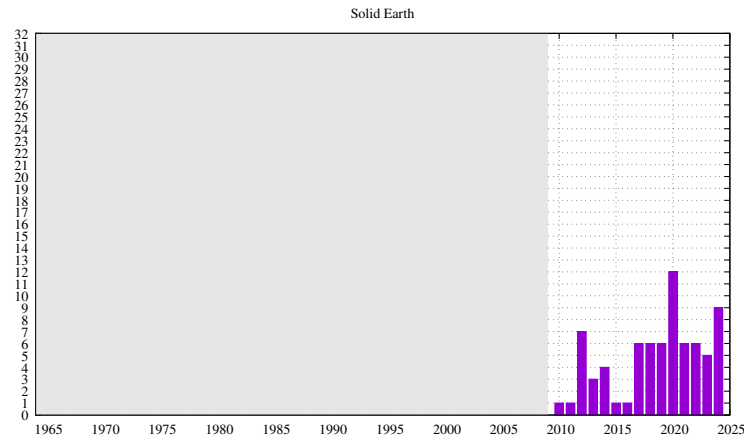
- **Geoscientific Model Development**[2] is a not-for-profit international scientific journal dedicated to the publication and public discussion of the description, development, and evaluation of numerical models of the Earth system and its components. It is published by EGU/Copernicus Publications and started in 2008.

---

[1] https://www.solid-earth.net/
[2] https://www.geoscientific-model-development.net/

36

Geosci. Model. Dev.
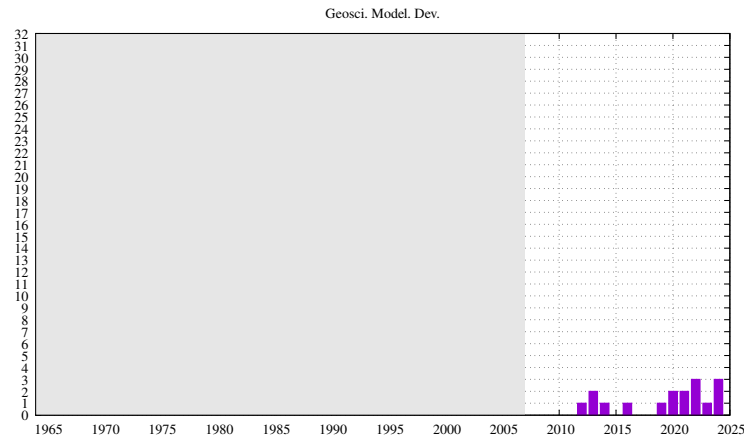
- **Physics of the Earth and Planetary Interiors**[3]. Launched in 1968 to fill the need for an international journal in the field of planetary physics, geodesy and geophysics, Physics of the Earth and Planetary Interiors has now grown to become important reading matter for all geophysicists. It is the only journal to be entirely devoted to the physical and chemical processes of planetary interiors. It is published by Elsevier.

Physics of the Earth and Planetary Interiors
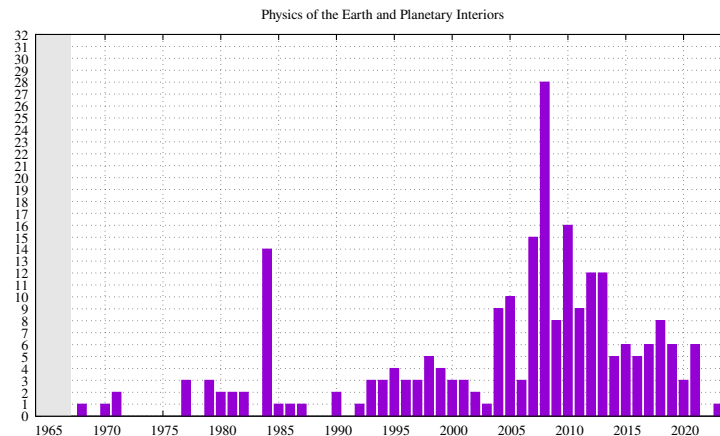
- **Geophysical Research Letters**[4] is a biweekly peer-reviewed scientific journal of geoscience published by the American Geophysical Union that was established in 1974. It is a gold open access journal that publishes high-impact, innovative, and timely communications-length articles on major advances spanning all of the major geoscience disciplines. Papers should have broad and immediate implications meriting rapid decisions and high visibility.

---

[3]https://www.sciencedirect.com/journal/physics-of-the-earth-and-planetary-interiors
[4]https://agupubs.onlinelibrary.wiley.com/journal/19448007

Geophysical Research Letters
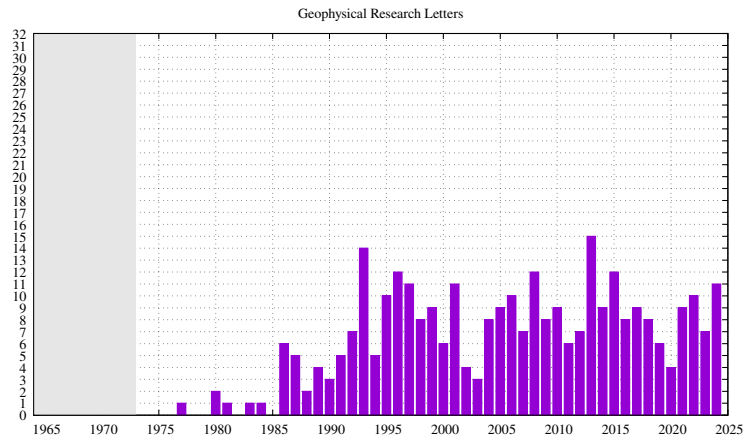
- **Geochemistry, Geophysics, Geosystems**[5] is an open access journal that publishes original research papers on Earth and planetary processes with a focus on understanding the Earth as a system. It is published by AGU and started in 2000.

Geochem. Geophys. Geosyst.
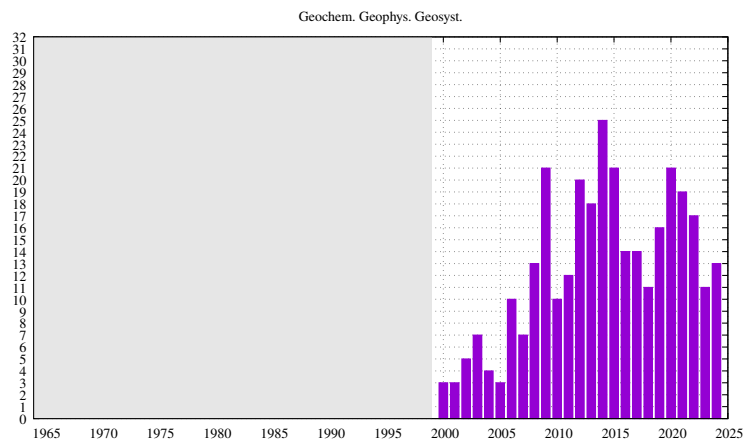
- **Earth and Planetary Science Letters**[6] is a leading journal for researchers across the entire Earth and planetary sciences community. It publishes concise, exciting, high-impact articles ("Letters") of broad interest. Its focus is on physical and chemical processes, the evolution and general properties of the Earth and planets - from their deep interiors to their atmospheres. It is published by Elsevier and started in 1966.

Earth Planet. Sci. Lett.
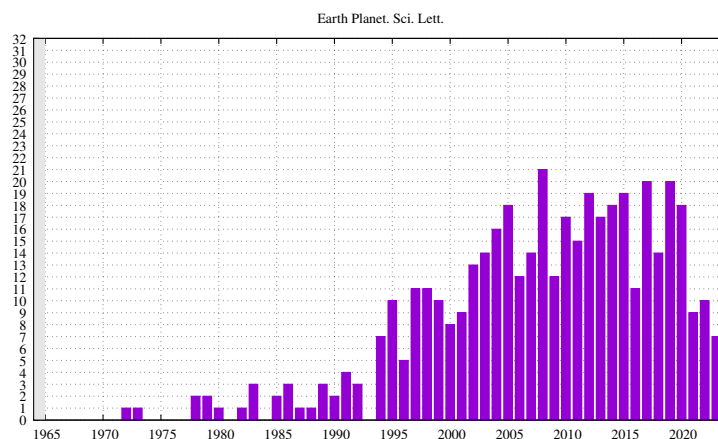
- **Journal of Geophysical Research** [7]

  The journal was originally founded under the name Terrestrial Magnetism by the American Geophysical Union's president Louis Agricola Bauer in 1896. It was renamed to Terrestrial Magnetism and Atmospheric Electricity in 1899 and in 1948 it acquired its current name. In 1980, three specialized sections were established: A: Space Physics, B: Solid Earth, and C: Oceans. Subsequently, further sections have been added: D: Atmospheres in 1984, E: Planets in 1991, F: Earth Surface in 2003, and G: Biogeosciences in 2005.

J. Geophys. Res.: Solid Earth

J. Geophys. Res.: Planets
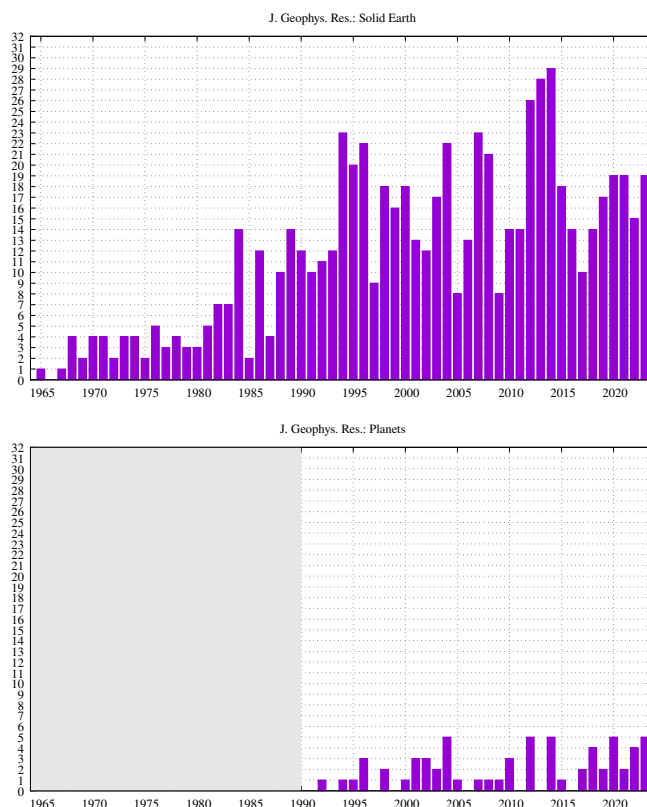
- **Geophysical Journal International**[8] Geophysical Journal International (GJI) is a monthly peer-reviewed scientific journal in the field of geophysics. It is published by Oxford University Press on behalf of two learned societies: the Royal Astronomical Society (RAS) and the Deutsche Geophysikalische Gesellschaft (German Geophysical Society, DGG), who select and peer-review the contents.

  GJI was formed in 1988 by the merger of three earlier geophysics journals (Geophysical Journal, Geophys. J. R. astr. Soc. and ?), the oldest of which was established in 1922. See Wiki[9] for a detailed history.

Geophysical Journal International
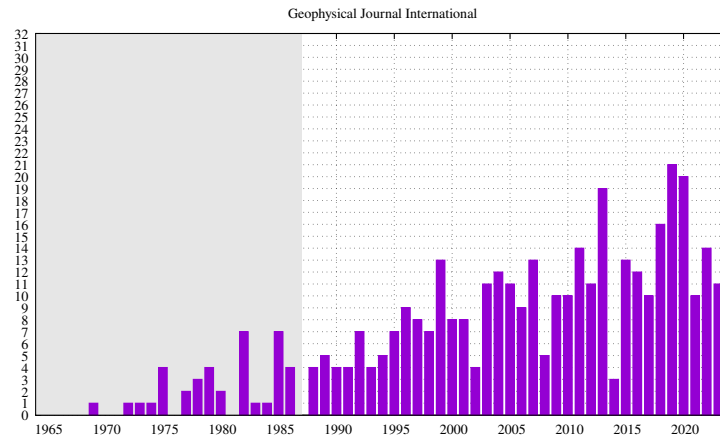
- **Geology**[10] The journal Geology publishes timely, innovative, and provocative articles relevant to its international audience, representing research from all fields of the geosciences. It is published by the GSA and started in 1973.
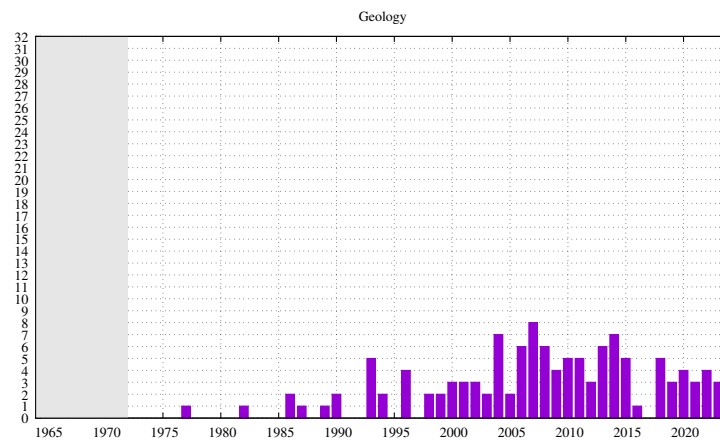
Geology

- **Tectonics**[11] Tectonics presents original research articles that describe and explain the evolution, structure, and deformation of Earth's lithosphere including across the range of geologic time. It is published by AGU and started in 1982.

Tectonics

---

[10]https://pubs.geoscienceworld.org/geology
[11]https://agupubs.onlinelibrary.wiley.com/journal/19449194
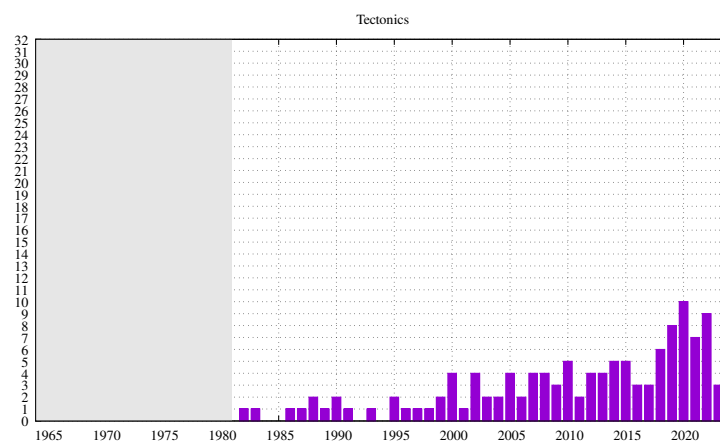
- **Tectonophysics**[12]: The prime focus of Tectonophysics will be high-impact original research and reviews in the fields of kinematics, structure, composition, and dynamics of the solid earth at all scales. It is published by Elsevier and started in 1964.

Tectonophysics
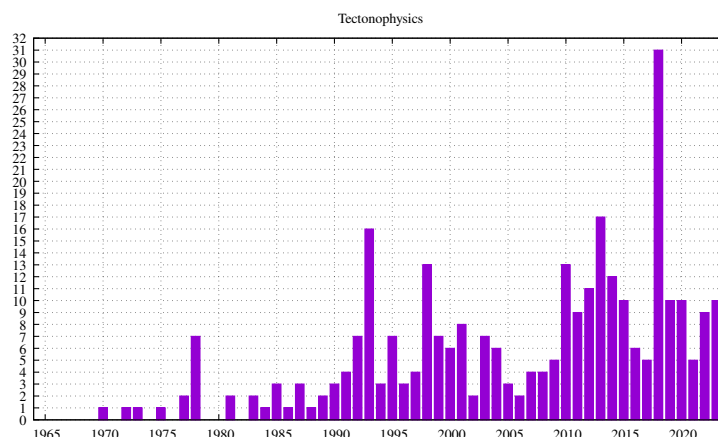
- **Computers & Geosciences**[13]: Computers & Geosciences publishes high impact, original research at the interface between Computer Sciences and Geosciences. Publications should apply modern computer science paradigms, whether computational or informatics-based, to address problems in the geosciences. It started in 1975.

Computers & Geosciences
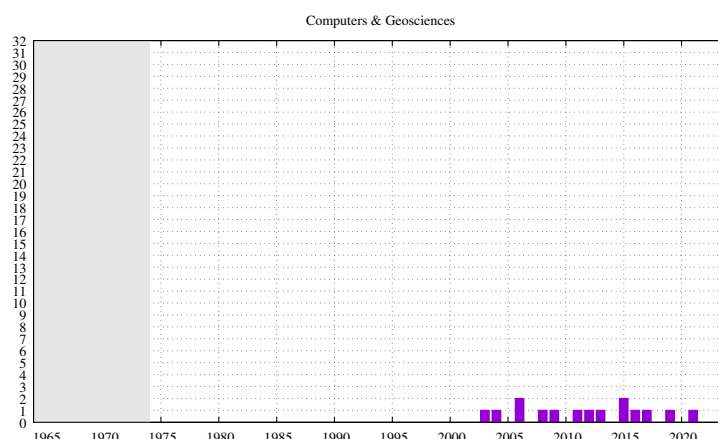
- **Computer Methods in Applied Mechanics and Engineering** [14]: Computer Methods in Applied Mechanics and Engineering was founded over five decades ago, providing a platform for the publication of papers in this important field of computational science and engineering. The range of appropriate contributions is very wide. It covers any type of computational method for the simulation of complex physical problems leading to the analysis and design of engineering products and systems. It started in June 1972.

---

[12]https://www.sciencedirect.com/journal/tectonophysics
[13]https://www.sciencedirect.com/journal/computers-and-geosciences
[14]https://www.sciencedirect.com/journal/computer-methods-in-applied-mechanics-and-engineering

Computer Methods in Applied Mechanics and Engineering
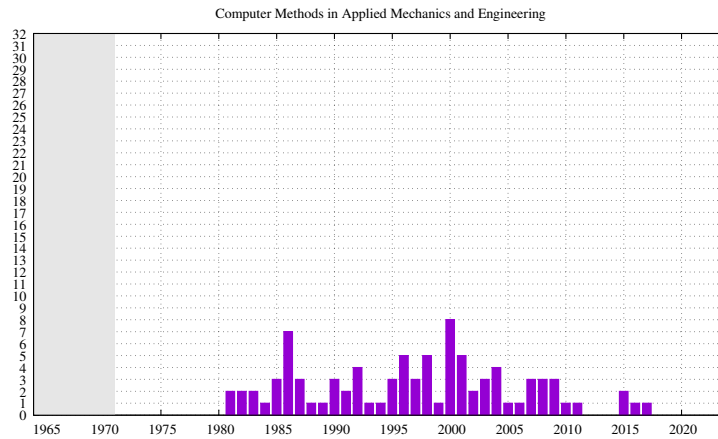
- **Earth and Planetary Physics** [15]:

  Earth and Planetary Physics (EPP) publishes original scholarship in Earth and planetary sciences, including solid-earth geophysics, atmospheric physics, space physics, and planetary sciences. Co-sponsored by the Chinese Geophysical Society, the Institute of Geology and Geophysics of the Chinese Academy of Sciences, and Science Press, EPP was hosted by AGU with Wiley in 2017-2021. Since 2022, EPP is operated independently. It started in November 2017.

Earth and Planetary Physics
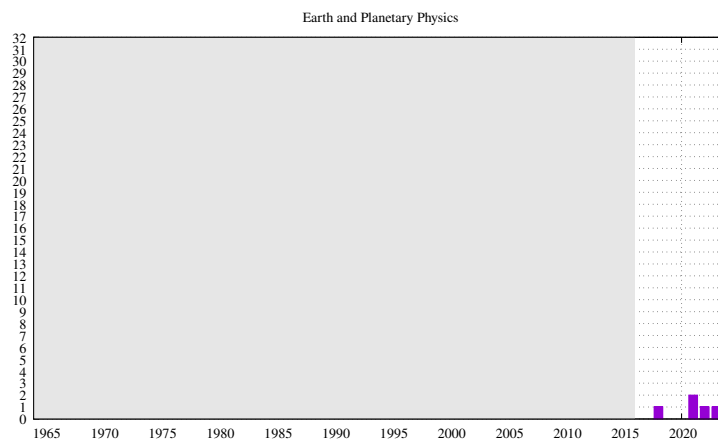
- **Frontiers in Earth Science**[16]: Frontiers in Earth Science is a multidisciplinary journal that welcomes research that enhances our understanding of the Earth's physical and chemical properties and processes, and aims to address major environmental challenges. I can't find out when it started.

---

[15]https://agupubs.onlinelibrary.wiley.com/journal/20963955,http://www.eppcgs.org/
[16]https://www.frontiersin.org/journals/earth-science

Frontiers in Earth Science
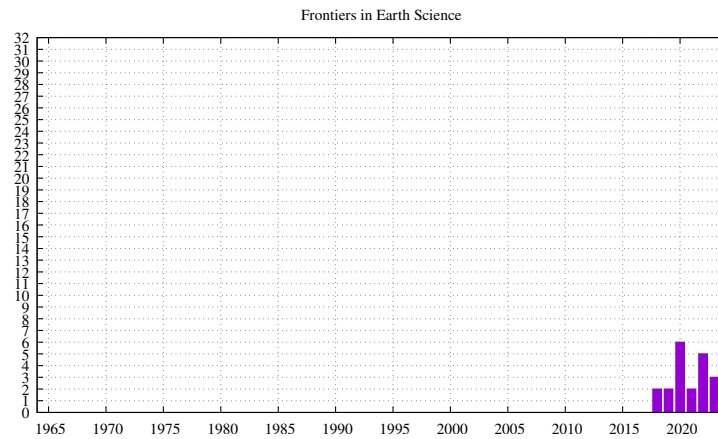
- **Gondwana Research**[17]: Gondwana Research is a peer-reviewed scientific journal with an "all earth science" scope and an emphasis on the origin and evolution of continents. It is part of the Elsevier group. It started in October 1997.

Gondwana Research
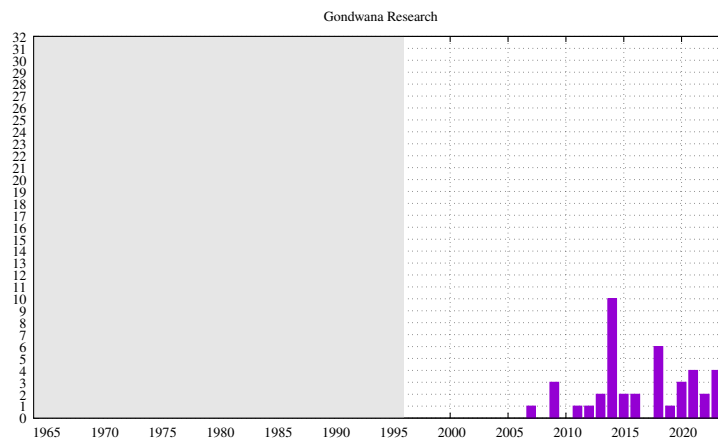
- **Icarus**[18]: Icarus is devoted to the publication of original contributions in the field of Solar System studies. Manuscripts reporting the results of new research - observational, experimental, or theoretical - concerning the astronomy, geology, meteorology, physics, chemistry, biology, and other scientific aspects of our Solar System or extrasolar systems are welcome. It started in 1962.

Icarus

---

[17]https://www.sciencedirect.com/journal/gondwana-research
[18]https://www.sciencedirect.com/journal/icarus
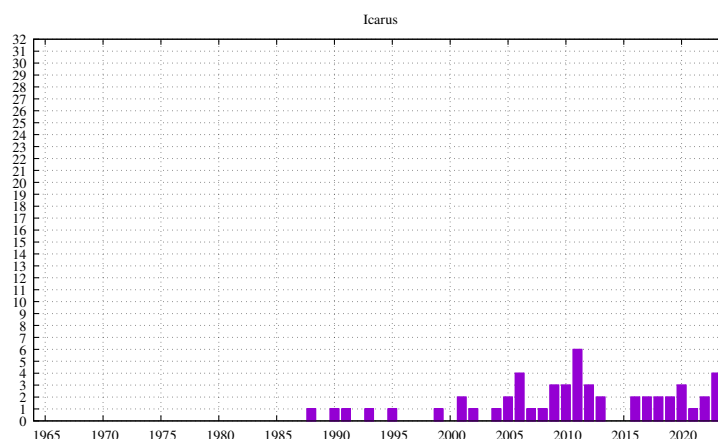
- **International Journal for Numerical Methods in Engineering** [19]:

  The International Journal for Numerical Methods in Engineering is known for pioneering techniques that help to solve critical engineering problems. This engineering mathematics journal covers a wide range of fundamental and emerging concepts in computational science and engineering. IJNME is the home for difference-making research, and we champion research on significant developments in all types of numerical methods, leading to global engineering solutions and progress.

Int. J. Num. Meth. Eng.
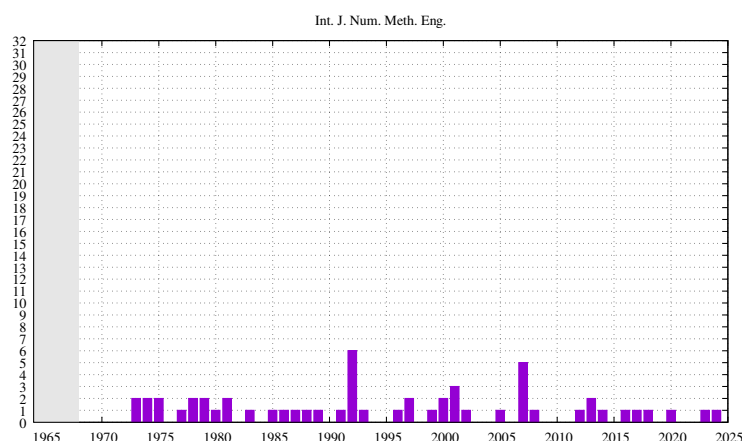
- **International Journal for Numerical Methods in Fluids** [20]: The International Journal for Numerical Methods in Fluids publishes refereed papers describing significant developments in computational methods that are applicable to scientific and engineering problems in fluid mechanics, fluid dynamics, micro and bio fluidics, and fluid-structure interaction. Numerical methods for solving ancillary equations, such as transport and advection and diffusion, are also relevant. The Editors encourage contributions in the areas of multi-physics, multi-disciplinary and multi-scale problems involving fluid subsystems, verification and validation, uncertainty quantification, and model reduction.

Int. J. Num. Meth. Fluids
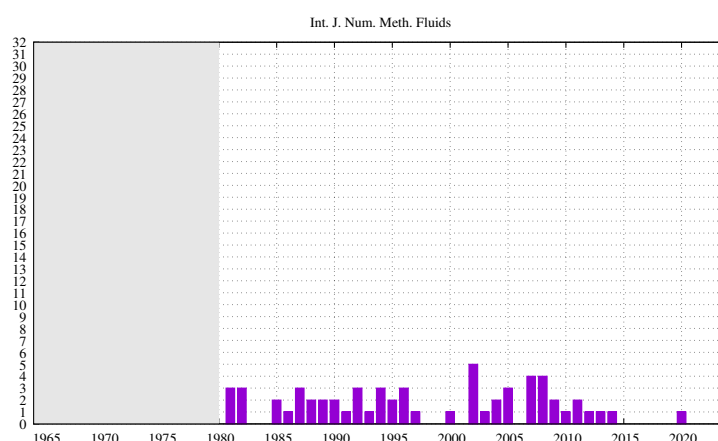
- **Journal of Geodynamics** [21]: The Journal of Geodynamics is an international and interdisciplinary forum for the publication of results and discussions of Solid Earth research covering the entire multitude of scientific methods aimed at understanding deep and shallow processes in

---

[19] https://onlinelibrary.wiley.com/journal/10970207
[20] https://onlinelibrary.wiley.com/journal/10970363
[21] https://www.sciencedirect.com/journal/journal-of-geodynamics

various tectonic settings and acting at different time- and length-scales. The Journal publishes results based on a multitude of research disciplines, including geophysics, geodesy, geochemistry, geology, and based on field observations, laboratory experiments, and numerical simulations.

Journal of Geodynamics
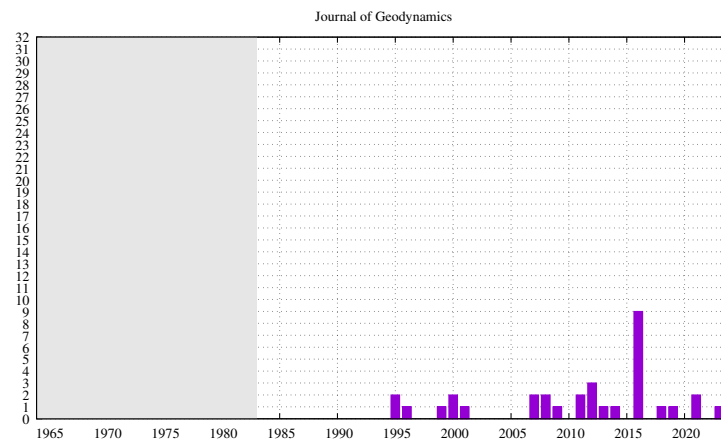
- **Journal of Open Source Software**[22]: Committed to publishing quality research software with zero article processing charges or subscription fees. I cannot find when it started.

Journal of Open Source Software
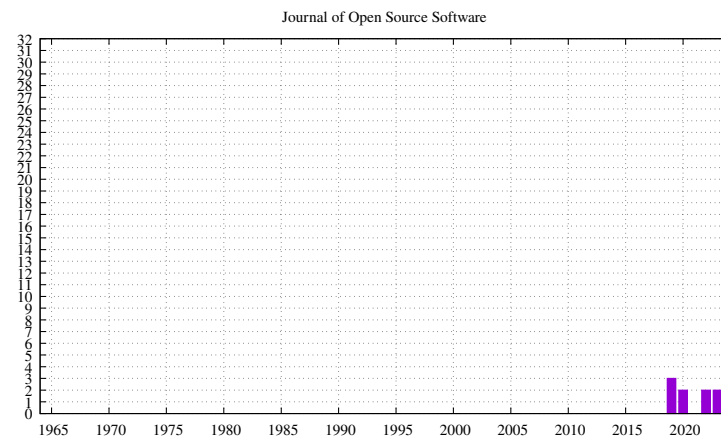
- **Journal of Structural Geology**[23]: The Journal of Structural Geology publishes process-oriented investigations about structural geology using appropriate combinations of analog and digital field data, seismic reflection data, satellite-derived data, geometric analysis, kinematic analysis, laboratory experiments, computer...

Journal of Structural Geology

---

[22]https://joss.theoj.org/
[23]https://www.sciencedirect.com/journal/journal-of-structural-geology
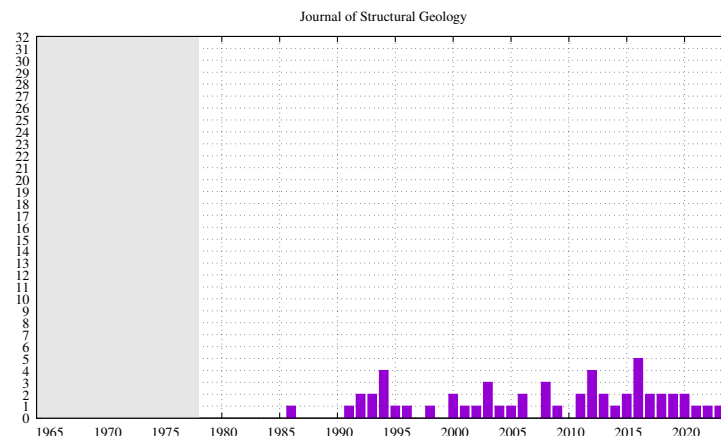
- **Pure & Applied Geophysics** [24]: Pure and Applied Geophysics is a long-standing journal focusing on the fields of solid Earth, atmospheric and oceanic sciences. It seems to have started in 1939.

Pure & Applied Geophysics
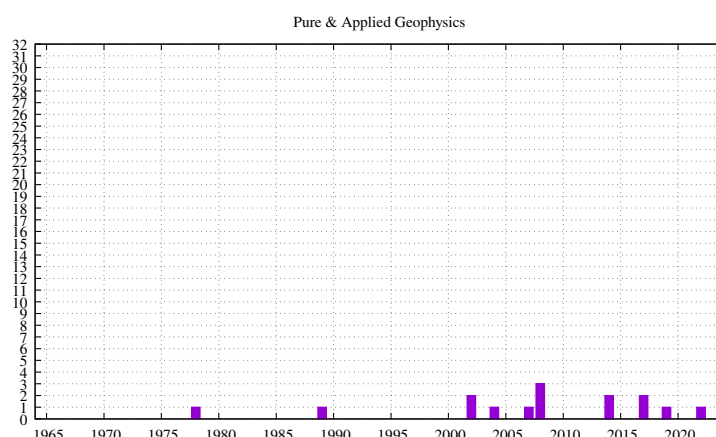
- **Proceedings of the National Academy of Sciences** [25]: It is the official journal of the National Academy of Sciences, published since 1915, and publishes original research, scientific reviews, commentaries, and letters.

Proceedings of the National Academy of Sciences
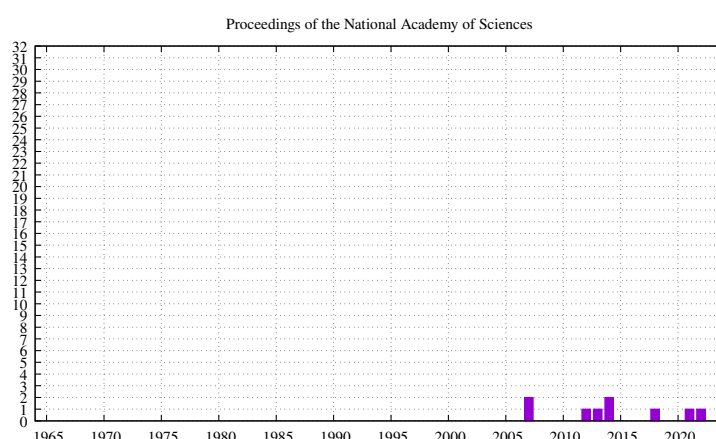
- **Terra Nova**[26]:

  Terra Nova is an earth science journal publishing short, innovative and provocative papers that expose new general principles and understanding, or challenge conventional wisdom. It started in 1989.

---

[24]https://link.springer.com/journal/24
[25]https://www.pnas.org/
[26]https://onlinelibrary.wiley.com/journal/13653121

Terra Nova
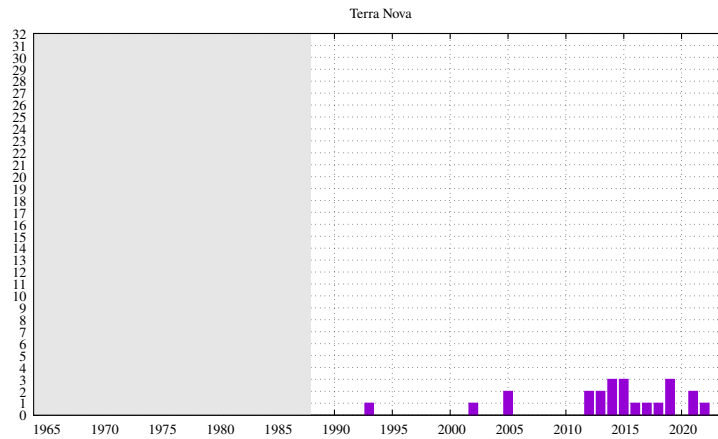
- **Scientific Reports**[27]: Scientific Reports is an open access journal publishing original research from across all areas of the natural sciences, psychology, medicine and engineering. It is is the 5th most-cited journal in the world, with more than 734,000 citations in 2023. The journal was established in 2011.

Scientific Reports
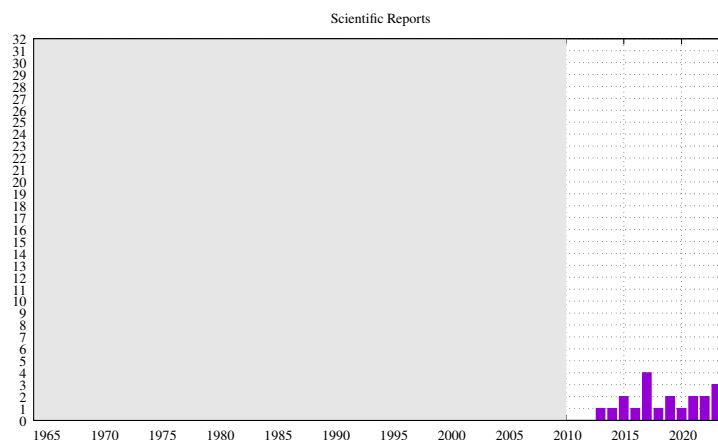
- **Nature(s)**[28]:

  Nature is a British weekly scientific journal founded and based in London, England. As a multidisciplinary publication, Nature features peer-reviewed research from a variety of academic disciplines, mainly in science and technology. It has core editorial offices across the United States, continental Europe, and Asia under the international scientific publishing company Springer Nature. Nature was one of the world's most cited scientific journals by the Science Edition of the 2022 Journal Citation Reports (with an ascribed impact factor of 50.5), making it one of the world's most-read and most prestigious academic journals.

  Nature Geoscience is a monthly peer-reviewed scientific journal published by the Nature Publishing Group. It was established in January 2008.

  Nature Communications is a peer-reviewed, open access, scientific journal published by Nature Portfolio since 2010. It is a multidisciplinary journal that covers the natural sciences, including physics, chemistry, earth sciences, medicine, and biology.

---

[27]https://www.nature.com/srep/

[28]https://www.nature.com/, https://www.nature.com/ngeo/, https://www.nature.com/ncomms/
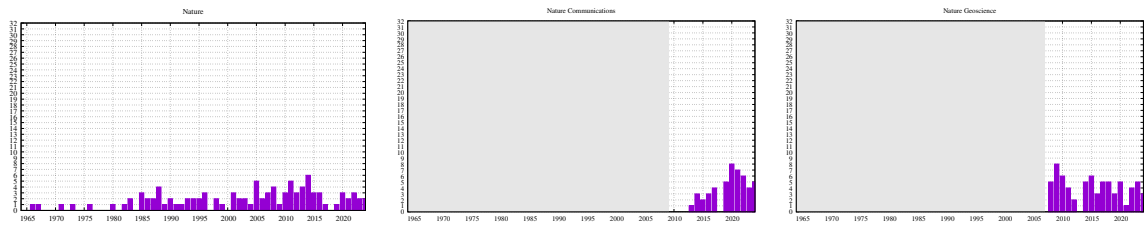
- **Geophysical & Astrophysical Fluid Dynamics**
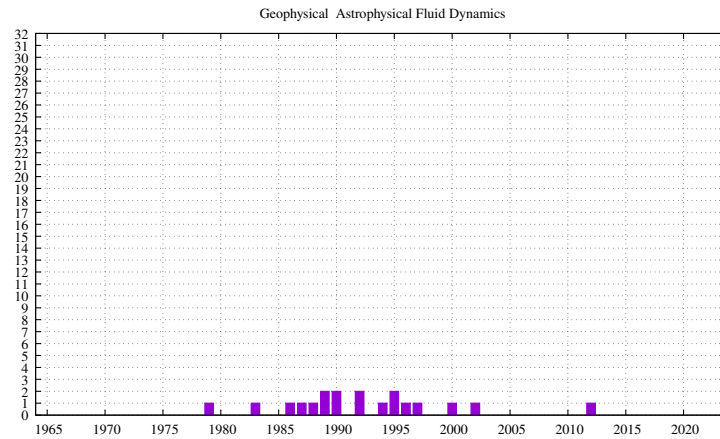
Chapter 3

Physics and a bit of mathematics

3.1 Some maths

3.1.1 About vectors

Remark. *In this document I have chosen to (when possible) use the notation $\vec{a}$ to denote a vector and $\boldsymbol{a}$ to denote a tensor/matrix. More often than not the same notation $\boldsymbol{a}$ is used for both in the literature.*

In mathematics, physics and engineering, a Euclidean vector or simply a vector is a geometric object that has magnitude (or length) and direction. Many algebraic operations on real numbers such as addition, subtraction, multiplication, and negation have close analogues for vectors.

Let $\vec{v}$ be a vector in 3D space. Its Euclidean norm (or magnitude) is given in a coordinate-free way by

$$|\vec{v}| := \sqrt{\vec{v} \cdot \vec{v}}$$

This definition makes use of the dot product, see next section. The Euclidean norm is also called the L_2-norm, or $2-$norm. It is also sometimes noted $|| \cdot ||_2$.

In Cartesian coordinates the vector $\vec{v}$ is given by

$$\vec{v} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = v_x \vec{e}_x + v_y \vec{e}_y + v_z \vec{e}_z \qquad \text{with} \qquad \vec{e}_x = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \vec{e}_y = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \vec{e}_z = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Its norm then simply writes

$$|\vec{v}| = \sqrt{v_x^2 + v_y^2 + v_z^2}$$

A unit vector is any vector with a length of one. A vector of arbitrary length can be divided by its length to create a unit vector. If $\vec{a}$ is a vector, the corresponding unit vector is often denoted

$$\vec{e}_a = \frac{\vec{a}}{|\vec{a}|}$$

3.1.2 dot products, cross products and dyadic products

The **dot product** (or sometimes called inner product, or even scalar product) of two vectors is denoted by $\vec{a} \cdot \vec{b}$ and is defined as:

$$\vec{a} \cdot \vec{b} = |\vec{a}|\, |\vec{b}|\, \cos\theta$$

where θ is the measure of the angle between $\vec{a}$ and b.

FIGURE

In Cartesian coordinates the dot product can also be defined as the sum of the products of the components of each vector as

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z$$

The dot product can also be interpreted as an answer to the question "how similar are vectors $\vec{a}$ and $\vec{b}$ in magnitude and direction?" Indeed, if $\vec{a} = \vec{b}$ then $\theta = 0$ and $\cos\theta = 1$, while if $\vec{a}$ is perpendicular to $\vec{b}$, then $\theta = \pi/2$, $\cos\theta = 0$ and $\vec{a} \cdot \vec{b} = 0$.

In Cartesian coordinates, we find that

$$\vec{v} \cdot \vec{e}_x = (v_x \vec{e}_x + v_y \vec{e}_y + v_z \vec{e}_z) \cdot \vec{e}_x = v_x \underbrace{\vec{e}_x \cdot \vec{e}_x}_{=1} + v_y \underbrace{\vec{e}_y \cdot \vec{e}_x}_{=0} + v_z \underbrace{\vec{e}_z \cdot \vec{e}_x}_{=0} = v_x$$

In this case the interpretation of $\vec{v} \cdot \vec{e}_x$ could be "how much of $\vec{v}$ is in the direction $\vec{e}_x$".

The **cross product** (also called the vector product or outer product) of two vectors is also a vector. It is denoted $\vec{a} \times \vec{b}$ and defined as

$$\vec{c} = \vec{a} \times \vec{b} = |\vec{a}|\, |\vec{b}|\, \sin\theta\, \vec{n}$$

where θ is the measure of the angle between $\vec{a}$ and b and and $\vec{n}$ is a unit vector perpendicular to both $\vec{a}$ and $\vec{b}$ which completes a right-handed system.

FIGURE

The norm of the cross product, say $|\vec{c}| = |\vec{a} \times \vec{b}|$, is actually the area of the parallelogram having $\vec{a}$ and $\vec{b}$ as sides.

Also note that $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$ (think about the direction of the normal vector in each case). In Cartesian coordinates the cross product can be written as

$$\vec{a} \times \vec{b} = (a_y b_z - a_z b_y)\vec{e}_x + (a_z b_x - a_x b_z)\vec{e}_y + (a_x b_y - a_y b_x)\vec{e}_z$$

Finally, let us look at the **dyadic product** of two vectors $\vec{a}$ and $\vec{b}$ which denoted by $\vec{a}\,\vec{b}^T$ (juxtaposed; no symbols, multiplication signs, crosses, dots, etc...). The result is a tensor:

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix}, \qquad \vec{a}\vec{b}^T = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} (b_x\ b_y\ b_z) = \begin{pmatrix} a_x b_x & a_x b_y & a_x b_z \\ a_y b_x & a_y b_y & a_y b_z \\ a_z b_x & a_z b_y & a_z b_z \end{pmatrix}$$

In conclusion the dot product yields a scalar, the cross product yields a vector and the dyadic product yields a tensor.

3.1.3 Rotation matrix

After much confusion, `https://mathworld.wolfram.com/RotationMatrix.html` is a source of clarity: one must be careful when speaking of 'rotation matrix'. Indeed, there are two possible conventions: rotation of the axes, and rotation of the object relative to fixed axes.

We consider in $\mathbb{R}^2$ the matrix $\boldsymbol{R}$ that rotates a given vector $\vec{v}$ by a counterclockwise angle θ in a fixed coordinate system. It writes

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

with $\vec{v}' = \boldsymbol{R} \cdot \vec{v}$.

On the other hand, consider the matrix that rotates the coordinate system through a counterclockwise angle θ. The coordinates of the fixed vector $\vec{v}$ in the rotated coordinate system are now given by a rotation matrix which is the transpose of the fixed-axis matrix and, as can be seen in the above diagram, is equivalent to rotating the vector by a counterclockwise angle of θ relative to a fixed set of axes, giving

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

In the following example we start from $\vec{v} = (2, 1)$. If we rotate the vector by 90°, the rotation matrix is given by

$$\boldsymbol{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

so that $\vec{v}' = (-1, 2)$. If we rotate the axis by 90°, the rotation matrix is given by

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and the coordinates of the resulting vector are $\vec{v}' = (1, -2)$.

(rotation_matrix.tex)
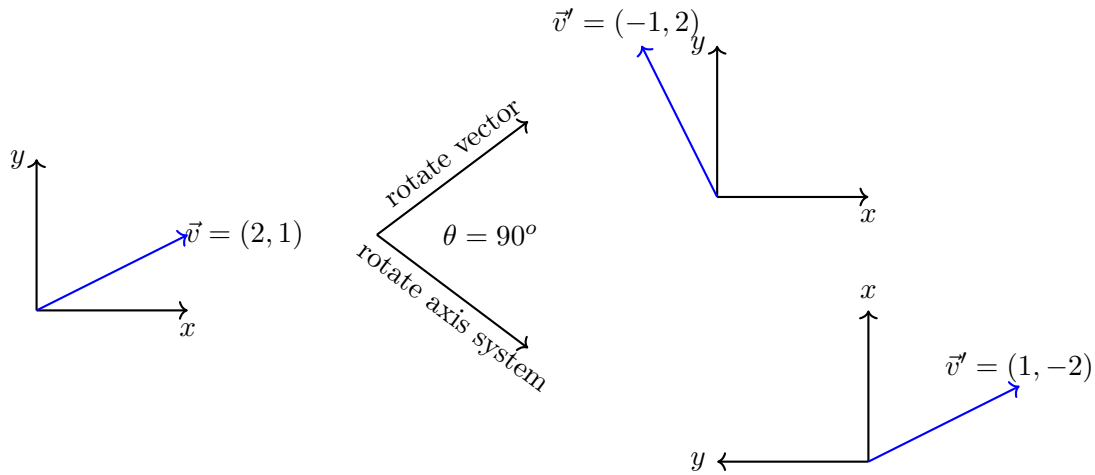
3.1.4 Inverse of a 2x2 matrix

Let us assume we wish to solve the system $\boldsymbol{A} \cdot \vec{X} = \vec{b}$, with $\vec{X} = (x, y)$. Then the solution is given by The solution is given by

$$x = \frac{1}{det(\boldsymbol{A})} \begin{vmatrix} b_1 & a_{21} \\ b_2 & a_{22} \end{vmatrix} \qquad y = \frac{1}{det(\boldsymbol{A})} \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}$$

3.1.5 Inverse of a 3x3 matrix

Let us consider the 3×3 matrix $\boldsymbol{M}$

$$\boldsymbol{M} = \begin{pmatrix} M_{xx} & M_{xy} & M_{xz} \\ M_{yx} & M_{yy} & M_{yz} \\ M_{zx} & M_{zy} & M_{zz} \end{pmatrix}$$

1. Find $det(\boldsymbol{M})$, the determinant of the Matrix $\boldsymbol{M}$. The determinant will usually show up in the denominator of the inverse. If the determinant is zero, the matrix won't have an inverse.

2. Find $\boldsymbol{M}^T$, the transpose of the matrix. Transposing means reflecting the matrix about the main diagonal.

$$\boldsymbol{M}^T = \begin{pmatrix} M_{xx} & M_{yx} & M_{zx} \\ M_{xy} & M_{yy} & M_{zy} \\ M_{xz} & M_{yz} & M_{zz} \end{pmatrix}$$

3. Find the determinant of each of the 2×2 minor matrices. For instance $\tilde{M}_{xx} = M_{yy}M_{zz} - M_{yz}M_{zy}$, or $\tilde{M}_{xz} = M_{xy}M_{yz} - M_{xz}M_{yy}$.

4. assemble the $\tilde{\boldsymbol{M}}$ matrix:

$$\tilde{\boldsymbol{M}} = \begin{pmatrix} +\tilde{M}_{xx} & -\tilde{M}_{xy} & +\tilde{M}_{xz} \\ -\tilde{M}_{yx} & +\tilde{M}_{yy} & -\tilde{M}_{yz} \\ +\tilde{M}_{zx} & -\tilde{M}_{zy} & +\tilde{M}_{zz} \end{pmatrix}$$

5. the inverse of $\boldsymbol{M}$ is then given by

$$\boldsymbol{M}^1 = \frac{1}{det(\boldsymbol{M})}\tilde{\boldsymbol{M}}$$

Another approach which of course is equivalent to the above is Cramer's rule. Let us assume we wish to solve the system $\boldsymbol{A} \cdot \vec{X} = \vec{b}$, with $\vec{X} = (x, y, z)$. Then the solution is given by

$$x = \frac{1}{det(\boldsymbol{M})} \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix} \qquad y = \frac{1}{det(\boldsymbol{M})} \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix} \qquad z = \frac{1}{det(\boldsymbol{M})} \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}$$

3.1.6 Symmetric matrices

Any *symmetric* matrix has only real eigenvalues, is always diagonalizable, and has orthogonal eigenvectors. A symmetric $N \times N$ real matrix $\boldsymbol{M}$ is said to be

- **positive definite** if $\vec{x} \cdot \boldsymbol{M} \cdot \vec{x} > 0$ for every non-zero vector $\vec{x}$ of n real numbers. All the eigenvalues of a Symmetric Positive Definite (SPD) matrix are positive. If A and B are positive definite, then so is A+B. The matrix inverse of a positive definite matrix is also positive definite. An SPD matrix has a unique Cholesky decomposition. In other words the matrix $\boldsymbol{M}$ is positive definite if and only if there exists a unique lower triangular matrix $\boldsymbol{L}$, with real and strictly positive diagonal elements, such that $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T$ (the product of a lower triangular matrix and its conjugate transpose). This factorization is called the Cholesky decomposition of $\boldsymbol{M}$.

- **positive semi-definite** if $\vec{x} \cdot \boldsymbol{M} \cdot \vec{x} \geq 0$

- **negative definite** if $\vec{x} \cdot \boldsymbol{M} \cdot \vec{x} < 0$

- **negative semi-definite** if $\vec{x} \cdot \boldsymbol{M} \cdot \vec{x} \leq 0$

The Stokes linear system

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{g} \end{pmatrix}$$

is **indefinite** (i.e. it has positive as well as negative eigenvalues).

A square matrix that is not invertible is called **singular** or degenerate. A square matrix is singular if and only if its determinant is 0. Singular matrices are rare in the sense that if you pick a random square matrix, it will almost surely not be singular.

3.1.7 Eigenvalues of positive definite matrix

Suppose our matrix $\boldsymbol{M}$ has eigenvalue λ.

If $\lambda = 0$, then there is some eigenvector $\vec{x}$ so that $\boldsymbol{M} \cdot \vec{x} = \lambda \vec{x} = \vec{0}$. But then $\vec{x}^T \cdot \boldsymbol{M} \cdot \vec{x} = 0$, and so $\mathcal{M}$ is not positive definite.

If $\lambda < 0$, then there is some eigenvector $\vec{x}$ so that $\boldsymbol{M} \cdot \vec{x} = \lambda \vec{x}$. But then $\vec{x}^T \cdot \boldsymbol{M} \cdot \vec{x} = \lambda |\vec{x}|^2$, which is negative since $|\vec{x}|^2 > 0$ and $\lambda < 0$. Thus $\boldsymbol{M}$ is not positive definite.

And so if $\boldsymbol{M}$ is positive definite, it only has positive eigenvalues.

3.1.8 Schur complement

From wiki. In linear algebra and the theory of matrices, the Schur complement of a matrix block (i.e., a submatrix within a larger matrix) is defined as follows. Suppose $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ are respectively $p \times p$, $p \times q$, $q \times p$ and $q \times q$ matrices, and $\mathbb{D}$ is invertible. Let

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$$

so that $\boldsymbol{M}$ is a $(p+q) \times (p+q)$ matrix. Then the Schur complement of the block $\boldsymbol{D}$ of the matrix $\boldsymbol{M}$ is the $p \times p$ matrix

$$\boldsymbol{S} = \boldsymbol{A} - \boldsymbol{B} \cdot \boldsymbol{D}^{-1} \cdot \boldsymbol{C}$$

Application to solving linear equations: The Schur complement arises naturally in solving a system of linear equations such as

$$\begin{aligned} \boldsymbol{A} \cdot \vec{x} + \boldsymbol{B} \cdot \vec{y} &= \vec{f} \\ \boldsymbol{C} \cdot \vec{x} + \boldsymbol{D} \cdot \vec{y} &= \vec{g} \end{aligned}$$

where $\vec{x}$, $\vec{f}$ are p-dimensional vectors, $\vec{y}$, $\vec{g}$ are q-dimensional vectors, and $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ are as above. Multiplying the bottom equation by $\boldsymbol{B} \cdot \boldsymbol{D}^{-1}$ and then subtracting from the top equation one obtains

$$(\boldsymbol{A} - \boldsymbol{B} \cdot \boldsymbol{D}^{-1} \cdot \boldsymbol{C}) \cdot \vec{x} = \vec{f} - \boldsymbol{B} \cdot \boldsymbol{D}^{-1} \cdot \vec{g}$$

Thus if one can invert $\boldsymbol{D}$ as well as the Schur complement of $\boldsymbol{D}$, one can solve for $\vec{x}$, and then by using the equation $\boldsymbol{C} \cdot \vec{x} + \boldsymbol{D} \cdot \vec{y} = \vec{g}$ one can solve for y. This reduces the problem of inverting a $(p+q) \times (p+q)$ matrix to that of inverting a $p \times p$ matrix and a $q \times q$ matrix. In practice one needs $\boldsymbol{D}$ to be well-conditioned in order for this algorithm to be numerically accurate.

Considering now the Stokes system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{v} \\ \vec{p} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{g} \end{pmatrix}$$

Factorising for $\vec{p}$ we end up with a **velocity-Schur complement**. Solving for $\vec{p}$ in the second equation and inserting the expression for $\vec{p}$ into the first equation we have

$$\mathbb{S}_v \cdot \vec{v} = \vec{f} \qquad \text{with} \qquad \mathbb{S}_v = \mathbb{K} + \mathbb{G} \cdot \mathbb{C}^{-1} \cdot \mathbb{G}^T$$

Factorising for $\vec{v}$ we get a **pressure-Schur complement**.

$$\mathbb{S}_p \cdot \vec{p} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \vec{f} \qquad \text{with} \qquad \mathbb{S}_p = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}$$

3.2   Units

Symbol	meaning	unit
t	Time	s
x, y, z	Cartesian coordinates	m
r, θ	Polar coordinates	m,-
r, θ, z	Cylindrical coordinates	m,-,m
r, θ, ϕ	Spherical coordinates	m,-,-
$\vec{v} = (u, v, w)$	velocity vector[1]	$\mathrm{m\,s^{-1}}$
$\vec{v} = (v_r, v_\theta, v_z)$	velocity vector[2]	$\mathrm{m\,s^{-1}}$
$\vec{v} = (v_r, v_\theta, v_\phi)$	velocity vector[3]	$\mathrm{m\,s^{-1}}$
$\vec{\upsilon} = (\upsilon_x, \upsilon_y, \upsilon_z)$	displacement vector	m
ρ	mass density	$\mathrm{kg\,m^{-3}}$
η	dynamic viscosity	Pa s
λ	penalty parameter	Pa s
T	temperature	K
$\vec{\nabla}$	gradient operator	$\mathrm{m^{-1}}$
$\vec{\nabla}\cdot$	divergence operator	$\mathrm{m^{-1}}$
p	pressure	Pa
$\dot{\varepsilon}(\vec{v})$	strain rate tensor	$\mathrm{s^{-1}}$
$\dot{\varepsilon}^d(\vec{v})$	deviatoric strain rate tensor	$\mathrm{s^{-1}}$
α	thermal expansion coefficient	$\mathrm{K^{-1}}$
k	thermal conductivity	$\mathrm{W\,m^{-1}\,K^{-1}}$
C_p	Heat capacity at constant pressure	$\mathrm{J\,kg^{-1}\,K^{-1}}$
H	intrinsic specific heat production	$\mathrm{W\,kg^{-1}}$
β_T	isothermal compressibility	$\mathrm{Pa^{-1}}$
$\boldsymbol{\tau}$	deviatoric stress tensor	Pa
$\boldsymbol{\sigma}$	full stress tensor	Pa
θ_L	Lode angle	-
λ	bulk modulus	Pa
μ	shear modulus	Pa
ν	Poisson ratio	-
E	Young's modulus	Pa

(1) Cartesian coordinates; (2) Cylindrical coordinates; (3) Spherical coordinates.

A quick note about units and LaTeX. This document relies on the `siunitx` package[2]. For instance, $\rho = 3300$ kg m^{-3} is obtained with

```
\rho = 3300~\si{\kg\per\cubic\metre}
```

or

```
\rho = \SI{3300}{\kg\per\cubic\metre}
```

Note that the `si` command can be used outside of the math environment.

3.3 Coordinate systems

coordinate_systems.tex

3.3.1 Cartesian coordinates

The unit vectors along the x, y and z axis are $\vec{e}_x$, $\vec{e}_y$ and $\vec{e}_z$ respectively.

(tikz_cartesian_coordinates.tex)

---
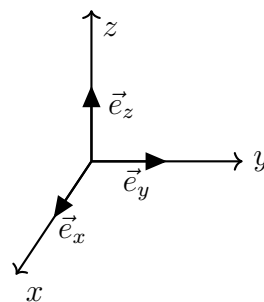
[1] https://en.wikipedia.org/wiki/International_System_of_Units
[2] https://ctan.org/pkg/siunitx

Any vector can then be written
$$\vec{V} = V_x \vec{e}_x + V_y \vec{e}_y + V_z \vec{e}_z$$

'How much of $\vec{V}$ is there in the $x-$direction' is obtained with $\vec{V} \cdot \vec{e}_x = V_x$. The gradient of a function f is
$$\vec{\nabla} f = \text{grad } f = \frac{\partial f}{\partial x} \, \vec{e}_x + \frac{\partial f}{\partial y} \, \vec{e}_y + \frac{\partial f}{\partial z} \, \vec{e}_z,$$

the divergence of a vector $\vec{V}$ is
$$\vec{\nabla} \cdot \vec{V} = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z}$$

and the Laplace operator of a function f is:
$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

Finally the path increment is
$$d\vec{r} = dx \, \vec{e}_x + dy \, \vec{e}_y + dz \, \vec{e}_z$$

and the volume element is
$$dV = dx \, dy \, dz.$$

3.3.2 Polar coordinates

We have $r > 0$ and $\theta = [0, 2\pi[$, defined in the $(x, y)-$plane.
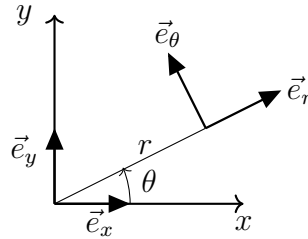
The relation between the unit vector in Cartesian and Polar/Cylindrical coordinates is given by:
$$\left(\begin{array}{c} \vec{e}_r \\ \vec{e}_\theta \end{array} \right) = \left(\begin{array}{cc} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{array} \right) \cdot \left(\begin{array}{c} \vec{e}_x \\ \vec{e}_y \end{array} \right)$$

which should be read:
$$\begin{array}{rcl} \vec{e}_r & = & \cos\theta \, \vec{e}_x + \sin\theta \, \vec{e}_y \\ \vec{e}_\theta & = & -\sin\theta \, \vec{e}_x + \cos\theta \, \vec{e}_y \end{array} \qquad (3.1)$$

Obviously for $\theta = 0$ we find $\vec{e}_r = \vec{e}_x$ and $\vec{e}_\theta = \vec{e}_y$, while for $\theta = \pi/2$ then $\vec{e}_r = \vec{e}_y$ and $\vec{e}_\theta = -\vec{e}_x$.

Note that this 2×2 matrix is a rotation matrix[3] corresponding to an angle $-\theta$. The inverse of this matrix always exists (we can always counter-rotate) and it then yields
$$\left(\begin{array}{c} \vec{e}_x \\ \vec{e}_y \end{array} \right) = \left(\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right) \cdot \left(\begin{array}{c} \vec{e}_r \\ \vec{e}_\theta \end{array} \right)$$

---

[3]https://en.wikipedia.org/wiki/Rotation_matrix

so that for any vector $\vec{V}$

$$
\begin{aligned}
\vec{V} &= V_x \vec{e}_x + V_y \vec{e}_y \\
&= V_x[(\cos\theta)\vec{e}_r - (\sin\theta)\vec{e}_\theta] + V_y[(\sin\theta)\vec{e}_r + (\cos\theta)\vec{e}_\theta] \\
&= [V_x(\cos\theta) + V_y(\sin\theta)]\vec{e}_r + [-V_x(\sin\theta) + V_y(\cos\theta)]\vec{e}_\theta \\
&= V_r \vec{e}_r + V_\theta \vec{e}_\theta
\end{aligned}
$$

with

$$
\begin{aligned}
V_r &= V_x \cos\theta + V_y \sin\theta \\
V_\theta &= -V_x \sin\theta + V_y \cos\theta
\end{aligned}
$$

Finally the path increment is

$$
d\vec{r} = dr\,\vec{e}_r + r\sin\theta d\theta\,\vec{e}_\theta
$$

and the volume element is

$$
dV = rdr\,d\theta.
$$

The gradient, divergence and Laplacian formulae are given in the following section about the cylindrical coordinates.

3.3.3   Cylindrical coordinates

Cylindrical coordinates
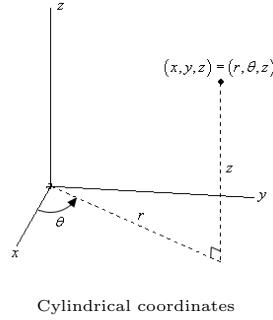
$$
\vec{V} = V_r\,\vec{e}_r + V_\theta\,\vec{e}_\theta + V_z\,\vec{e}_z
$$

We have

$$
\begin{aligned}
x &= r\,\cos\theta \\
y &= r\,\sin\theta \\
r &= \sqrt{x^2 + y^2}
\end{aligned}
$$

Let $f(r,\theta)$ be a function of the spatial coordinates. Its gradient is then

$$
\vec{\nabla}f = \frac{\partial f}{\partial r}\,\vec{e}_r + \frac{1}{r}\frac{\partial f}{\partial\theta}\,\vec{e}_\theta + \frac{\partial f}{\partial r}\,\vec{e}_z
$$

The divergence of a vector field $\vec{V}$ is

$$
\vec{\nabla}\cdot\vec{V} = \frac{1}{r}\frac{\partial}{\partial r}(rV_r) + \frac{1}{r}\frac{\partial V_\theta}{\partial\theta} + \frac{\partial V_z}{\partial z}
$$

and the Laplacian of f is

$$
\Delta f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2} + \frac{\partial^2 f}{\partial z^2}
$$

Finally the path increment is
$$d\vec{r} = dr\ \vec{e}_r + r\sin\theta d\theta\ \vec{e}_\theta + dz\ \vec{e}_z$$

and the volume element is
$$dV = rdr\ d\theta\ dz$$

Remark. *Cylindrical coordinates can also be denoted by (ρ, θ), (r, ϕ) or even (ρ, ϕ). They are some-times called "cylindrical polar coordinates" or "polar cylindrical coordinates".*

The divergence of the second order tensor field $\boldsymbol{S}$ in cylindrical polar coordinates is given by

$$\vec{\nabla} \cdot \boldsymbol{S} = \frac{\partial S_{rr}}{\partial r}\ \vec{e}_r + \frac{\partial S_{r\theta}}{\partial r}\ \vec{e}_\theta + \frac{\partial S_{rz}}{\partial r}\ \vec{e}_z$$
$$+ \frac{1}{r}\left[\frac{\partial S_{r\theta}}{\partial \theta} + (S_{rr} - S_{\theta\theta})\right]\ \vec{e}_r + \frac{1}{r}\left[\frac{\partial S_{\theta\theta}}{\partial \theta} + (S_{r\theta} + S_{\theta r})\right]\ \vec{e}_\theta + \frac{1}{r}\left[\frac{\partial S_{\theta z}}{\partial \theta} + S_{rz}\right]\ \vec{e}_z$$
$$+ \frac{\partial S_{zr}}{\partial z}\ \vec{e}_r + \frac{\partial S_{z\theta}}{\partial z}\ \vec{e}_\theta + \frac{\partial S_{zz}}{\partial z}\ \vec{e}_z$$

In the case of polar coordinates then all quantities featuring z (or ∂_z) are removed:

$$\vec{\nabla} \cdot \boldsymbol{S} = \frac{\partial S_{rr}}{\partial r}\ \vec{e}_r + \frac{\partial S_{r\theta}}{\partial r}\ \vec{e}_\theta + \frac{1}{r}\left[\frac{\partial S_{r\theta}}{\partial \theta} + (S_{rr} - S_{\theta\theta})\right]\ \vec{e}_r + \frac{1}{r}\left[\frac{\partial S_{\theta\theta}}{\partial \theta} + (S_{r\theta} + S_{\theta r})\right]\ \vec{e}_\theta$$
$$= \left[\frac{\partial S_{rr}}{\partial r} + \frac{1}{r}\frac{\partial S_{r\theta}}{\partial \theta} + \frac{1}{r}(S_{rr} - S_{\theta\theta})\right]\ \vec{e}_r + \left[\frac{\partial S_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial S_{\theta\theta}}{\partial \theta} + \frac{2}{r}S_{r\theta}\right]\ \vec{e}_\theta \qquad (3.2)$$

where we have assumed that the tensor $\boldsymbol{S}$ is symmetric (i.e. $S_{r\theta} = S_{\theta r}$).

3.3.4 Spherical coordinates

On the following figure are represented the three Cartesian axis, a point and its spherical coordinates r, θ, ϕ:

Spherical coordinates as commonly used in physics:

polar angle θ, and azimuthal angle ϕ.
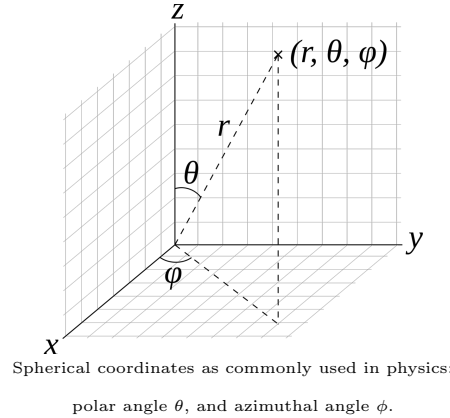
In this case $\theta \in [0 : \pi]$ and $\phi \in] - \pi : \pi]$ and we have the following relationships:

$$r = \sqrt{x^2 + y^2 + z^2} \qquad (3.3)$$
$$\theta = \arccos(z/r) \qquad (3.4)$$
$$\phi = \arctan(y/x) \qquad (3.5)$$
$$x = r\sin\theta\cos\phi \qquad (3.6)$$
$$y = r\sin\theta\sin\phi \qquad (3.7)$$
$$z = r\cos\theta \qquad (3.8)$$

The inverse tangent used to compute ϕ must be suitably defined, taking into account the correct quadrant of (x, y), which is why the atan2 intrinsic function is used in FORTRAN for example. This is often written as follows:

$$\theta = \arctan\left(\sqrt{x^2 + y^2}, z\right) \tag{3.9}$$

$$\phi = \arctan(y, x) \tag{3.10}$$

where we formally take advantage of the two argument arctan function to eliminate quadrant confusion.

The path increment is expressed as:

$$d\vec{r} = dr \ \vec{e}_r + r d\theta \ \vec{e}_\theta + r \sin\theta d\phi \ \vec{e}_\phi \tag{3.11}$$

The gradient of a function $f(r, \theta, \phi)$ is

$$\vec{\nabla} f = \frac{\partial f}{\partial r} \ \vec{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta} \ \vec{e}_\theta + \frac{1}{r \ \sin\theta}\frac{\partial f}{\partial \phi} \ \vec{e}_\phi \tag{3.12}$$

The divergence of a vector $\vec{V}$ is

$$\vec{\nabla} \cdot \vec{V} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 V_r\right) + \frac{1}{r \sin\theta}\frac{\partial}{\partial \theta}(V_\theta \sin\theta) + \frac{1}{r \sin\theta}\frac{\partial V_\phi}{\partial \phi} = 0 \tag{3.13}$$

The Laplacian of function f is given by:

$$\Delta f = \vec{\nabla}^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 f}{\partial \phi^2} \tag{3.14}$$

In geography one uses latitude and longitude, represented hereunder:
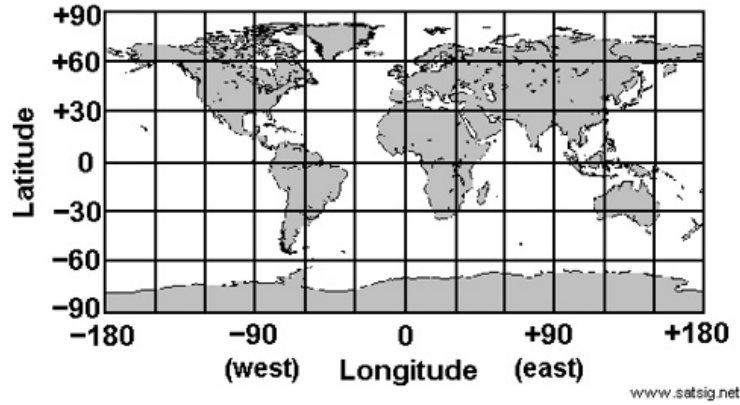
- Latitude $\in [-90 : 90]$, or $\in [-\pi/2 : \pi/2]$

- Longitude $\in]-180 : 180]$, or $\in]-\pi : \pi]$

Since the colatitude is the complementary angle of the latitude, i.e. the difference between 90 and the latitude, where southern latitudes are denoted with a minus sign, θ as shown above is actually is the colatitude. The colatitude is shown in red on the following figure:
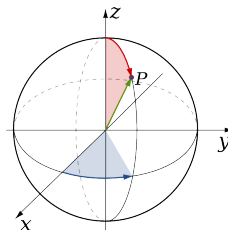
The volume of a sphere of radius R is easily obtained by computing

$$
\begin{aligned}
V_{sphere} &= \iiint_{sphere} dV \\
&= \int_0^R r^2 dr \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \\
&= \frac{1}{3}R^3 \cdot 2 \cdot 2\pi \\
&= \frac{4}{3}\pi R^3
\end{aligned}
\tag{3.15}
$$

The volume of a spherical shell of inner radius R_i and outer radius R_o is equally easily obtained by computing

$$
\begin{aligned}
V_{shell} &= \iiint_{shell} dV \\
&= \int_{R_i}^{R_o} r^2 dr \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \\
&= \frac{1}{3}(R_o^3 - R_i^3) \cdot 2 \cdot 2\pi \\
&= \frac{4}{3}\pi(R_o^3 - R_i^3)
\end{aligned}
\tag{3.16}
$$

The spherical unit vectors are related to the Cartesian unit vectors by:

$$
\begin{pmatrix} \vec{e}_r \\ \vec{e}_\theta \\ \vec{e}_\phi \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{pmatrix} \begin{pmatrix} \vec{e}_x \\ \vec{e}_y \\ \vec{e}_z \end{pmatrix}
$$

and the Cartesian unit vectors are related to the spherical unit vectors by

$$
\begin{pmatrix} \vec{e}_x \\ \vec{e}_y \\ \vec{e}_z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{pmatrix} \begin{pmatrix} \vec{e}_r \\ \vec{e}_\theta \\ \vec{e}_\phi \end{pmatrix}
$$

Finally, the velocity vector $\vec{v}$ then becomes

$$
\begin{aligned}
\vec{v} &= u\,\vec{e}_x + v\,\vec{e}_y + w\,\vec{e}_z \\
&= u\left(\sin\theta\cos\phi\,\vec{e}_r + \cos\theta\cos\phi\,\vec{e}_\theta - \sin\phi\,\vec{e}_\phi\right) \\
&+ v\left(\sin\theta\sin\phi\,\vec{e}_r + \cos\theta\sin\phi\,\vec{e}_\theta + \cos\phi\,\vec{e}_\phi\right) \\
&+ w\left(\cos\theta\,\vec{e}_r - \sin\theta\,\vec{e}_\theta\right) \\
&= v_r\,\vec{e}_r + v_\theta\,\vec{e}_\theta + v_\phi\,\vec{e}_\phi
\end{aligned}
\tag{3.17}
$$

with

$$
\begin{aligned}
v_r &= u\sin\theta\cos\phi + v\sin\theta\sin\phi + w\cos\theta \\
v_\theta &= u\cos\theta\cos\phi + v\cos\theta\sin\phi - w\sin\theta \\
v_\phi &= -u\sin\phi + v\cos\phi
\end{aligned}
\tag{3.18}
$$

3.3.5 Converting tensors between Cartesian and Cylindrical bases

$$
\boldsymbol{T}_{Cyl} = \begin{pmatrix} T_{rr} & T_{r\theta} & T_{rz} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta z} \\ T_{zr} & T_{z\theta} & T_{zz} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}
$$

$$
\boldsymbol{T}_{Cart} = \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} T_{rr} & T_{r\theta} & T_{rz} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta z} \\ T_{zr} & T_{z\theta} & T_{zz} \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}
$$

3.3.6 Converting tensors between Cartesian and Spherical bases

Let $\boldsymbol{T}$ be a tensor

$$
\boldsymbol{T} = \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \qquad \boldsymbol{T} = \begin{pmatrix} T_{rr} & T_{r\theta} & T_{r\phi} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta\phi} \\ T_{\phi r} & T_{\phi\theta} & T_{\phi\phi} \end{pmatrix}
$$

in the Cartesian basis (left) and the spherical basis (right).

The two sets of components are related by

$$
\begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} = \begin{pmatrix} \sin\theta\,\cos\phi & \cos\theta\,\cos\phi & -\sin\phi \\ \sin\theta\,\sin\phi & \cos\theta\,\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{pmatrix} \cdot \begin{pmatrix} T_{rr} & T_{r\theta} & T_{r\phi} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta\phi} \\ T_{\phi r} & T_{\phi\theta} & T_{\phi\phi} \end{pmatrix} \cdot \begin{pmatrix} \sin\theta\,\cos\phi & \sin\theta\,\sin\phi \\ \cos\theta\,\cos\phi & \cos\theta\,\sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}
$$

or

$$
\begin{pmatrix} T_{rr} & T_{r\theta} & T_{r\phi} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta\phi} \\ T_{\phi r} & T_{\phi\theta} & T_{\phi\phi} \end{pmatrix} = \begin{pmatrix} \sin\theta\,\cos\phi & \sin\theta\,\sin\phi & \cos\theta \\ \cos\theta\,\cos\phi & \cos\theta\,\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{pmatrix} \cdot \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \cdot \begin{pmatrix} \sin\theta\,\cos\phi & \cos\theta\,\cos\phi \\ \sin\theta\,\sin\phi & \cos\theta\,\sin\phi \\ \cos\theta & -\sin\theta \end{pmatrix}
$$

If we now assume that the tensor $\boldsymbol{T}$ is symmetric (e.g. stress tensor, strain rate tensor), then there are only 6 independent terms.

3.4 A continuum mechanics primer

Contains contributions by W. Spakman - Continuum mechanics course syllabus

3.4.1 Forces

In continuum mechanics we make a distinction between two broad classes of forces:

- Body forces defined as force per unit volume ($\mathrm{N\,m^{-3}}$): gravity, electro-magnetic forces

- Tractions: Surface forces defined as force per unit surface area ($\mathrm{N\,m^{-2}}$): Contact forces, elastic forces per unit area, internal flow friction, pressure, ...
  A traction is the surface average of all atomic forces exerted by atoms on the one side on atoms on the other side of the surface. For real-Earth processes, internal tractions are ultimately caused by the body forces, usually gravity.

  Existing mantle flow(i.e. flow that is forced elsewhere) can exert tractions (shear stresses) on the subducting slab or for instance at the base of lithosphere plates. In HPT-laboratory experiments external tractions (pressure, shear traction) are applied to a rock sample, which cause internal tractions to balance the exerted forces.

-blue slab: slab morphology without external mantle inflow
-yellow slab morphology after 25 My of trench-parallel mantle inflow
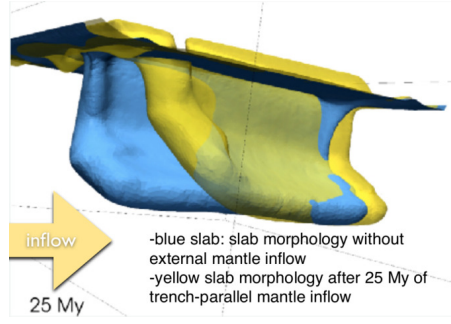
3.4.2 Stress tensor and tractions

The Cauchy stress tensor[4] consists of nine components σ_{ij} that completely define the state of stress at a point inside a material. The tensor relates a unit-length direction vector $\vec{n}$ to the so-called 'stress vector' (most commonly called 'traction') $\vec{t}(\vec{n})$ across an imaginary surface perpendicular to $\vec{n}$:

$$\vec{t}(\vec{n}) = \boldsymbol{\sigma} \cdot \vec{n}$$
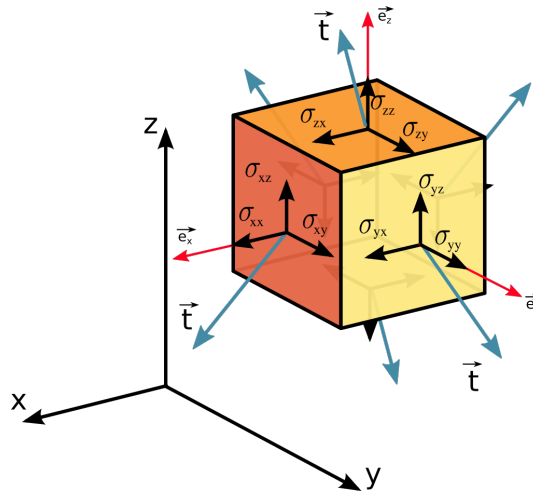
Modified from original file on Wikipedia[5]

With respect to an orthonormal basis $\{\vec{e}_x, \vec{e}_y, \vec{e}_z\}$, the Cauchy stress tensor is given by:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \tag{3.19}$$

The three diagonal elements are called normal stresses while the off-diagonal terms are called shear stresses.

One can easily prove (see for instance Section 3.3.6 of [522]) that the balance of angular momentum leads reduces to the statement that the Cauchy stress tensor is symmetric, i.e. $\boldsymbol{\sigma} = \boldsymbol{\sigma}^T$. Therefore, the stress state of the medium at any point and instant can be specified by only six independent parameters, rather than nine:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \quad \text{or sometimes} \quad \boldsymbol{\sigma} = \begin{pmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \sigma_y & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \sigma_z \end{pmatrix} \tag{3.20}$$

---

[4]https://en.wikipedia.org/wiki/Cauchy_stress_tensor
[5]https://commons.wikimedia.org/wiki/File:Components_stress_tensor_cartesian.svg

where the elements σ_x, σ_y, σ_z are called the orthogonal normal stresses (relative to the chosen coordinate system), and τ_{xy}, τ_{xz}, τ_{yz} the orthogonal shear stresses. The left form is preferred in this document. As seen above, the SI units of both stress tensor and traction are $\mathrm{N\,m^{-2}}$.

In Cylindrical coordinates the stress tensor components are given by:

$$\sigma_{rr} = -p + 2\eta \frac{\partial \mathbf{v}_r}{\partial r} \tag{3.21}$$

$$\sigma_{\theta\theta} = -p + 2\eta \left(\frac{1}{r} \frac{\partial \mathbf{v}_\theta}{\partial \theta} + \frac{\mathbf{v}_r}{r} \right) \tag{3.22}$$

$$\sigma_{zz} = -p + 2\eta \frac{\partial \mathbf{v}_z}{\partial z} \tag{3.23}$$

$$\sigma_{r\theta} = \eta \left(\frac{1}{r} \frac{\partial \mathbf{v}_r}{\partial \theta} + \frac{\partial \mathbf{v}_\theta}{\partial r} - \frac{\mathbf{v}_\theta}{r} \right) \tag{3.24}$$

$$\sigma_{rz} = \eta \left(\frac{\partial \mathbf{v}_r}{\partial z} + \frac{\partial \mathbf{v}_z}{\partial r} \right) \tag{3.25}$$

$$\sigma_{\theta z} = \eta \left(\frac{1}{r} \frac{\partial \mathbf{v}_z}{\partial \theta} + \frac{\partial \mathbf{v}_\theta}{\partial z} \right) \tag{3.26}$$

In Spherical coordinates the stress tensor components are given by:

$$\sigma_{rr} = -p + 2\eta \frac{\partial \mathbf{v}_r}{\partial r} \tag{3.27}$$

$$\sigma_{\theta\theta} = -p + 2\eta \left(\frac{1}{r} \frac{\partial \mathbf{v}_\theta}{\partial \theta} + \frac{\mathbf{v}_r}{r} \right) \tag{3.28}$$

$$\sigma_{\phi\phi} = -p + 2\eta \left(\frac{1}{r\sin\theta} \frac{\partial \mathbf{v}_\phi}{\partial \phi} + \frac{\mathbf{v}_r}{r} + \frac{\mathbf{v}_\theta \cot\theta}{r} \right) \tag{3.29}$$

$$\sigma_{r\theta} = \eta \left(r \frac{\partial}{\partial r} \frac{\mathbf{v}_\theta}{r} + \frac{1}{r} \frac{\partial \mathbf{v}_r}{\partial \theta} \right) \tag{3.30}$$

$$\sigma_{r\phi} = \eta \left(\frac{1}{r\sin\theta} \frac{\partial \mathbf{v}_r}{\partial \phi} + r \frac{\partial}{\partial r} \frac{\mathbf{v}_\phi}{r} \right) \tag{3.31}$$

$$\sigma_{\theta\phi} = \eta \left(\frac{1}{r\sin\theta} \frac{\partial \mathbf{v}_\theta}{\partial \phi} + \frac{\sin\theta}{r} \frac{\partial}{\partial \theta} \frac{\mathbf{v}_\phi}{\sin\theta} \right) \tag{3.32}$$

3.4.3 Strain rate and spin tensor

The velocity gradient $\boldsymbol{L}$ is given in Cartesian coordinates by:

$$\boldsymbol{L}(\vec{\mathbf{v}}) = \vec{\nabla}\vec{\mathbf{v}} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} & \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} & \frac{\partial v}{\partial z} & \frac{\partial w}{\partial z} \end{pmatrix} \tag{3.33}$$

It can be decomposed into its symmetric and skew-symmetric parts according to:

$$\vec{\nabla}\vec{\mathbf{v}} = (\vec{\nabla}\vec{\mathbf{v}})^s + (\vec{\nabla}\vec{\mathbf{v}})^w = \dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}}) + \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}) \tag{3.34}$$

The symmetric part is called the strain rate (or rate of deformation)[6]:

$$\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathrm{v}} + (\vec{\nabla}\vec{\mathrm{v}})^T\right) \tag{3.35}$$

The skew-symmetric tensor is called spin tensor (or vorticity tensor):

$$\dot{\boldsymbol{\omega}}(\vec{\mathrm{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathrm{v}} - (\vec{\nabla}\vec{\mathrm{v}})^T\right) \tag{3.36}$$

Remark. *In the mathematical literature a different notation for the strain rate tensor is often used, i.e. $\boldsymbol{D}(\vec{\mathrm{v}})$ - or simply $\boldsymbol{D}$, such as for instance in Fullsack (1995) [452].*

3.5 Viscous Newtonian rheology

The relationship between velocity-related stresses and velocity derivatives is such that the total stress tensor has the form [79]

$$\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{A} : \dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) \tag{3.37}$$

where p is the thermodynamic pressure which is a function of the density ρ and the temperature T (an equation of state is then needed) and $\boldsymbol{A}$ is the fourth-rank stiffness tensor.

Since both the stress and the strain tensors are symmetric and for isotropic fluids we have (see Malvern [879])

$$\boldsymbol{A} : \dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) = \lambda(\vec{\nabla}\cdot\vec{\mathrm{v}})\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) \tag{3.38}$$

where λ is the bulk viscosity and η is the dynamic viscosity[7]. The stress tensor is then

$$\boldsymbol{\sigma} = (-p + \lambda(\vec{\nabla}\cdot\vec{\mathrm{v}}))\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) \tag{3.39}$$

By writing

$$\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) = \frac{1}{3}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}))\mathbf{1} + \dot{\boldsymbol{\varepsilon}}^d(\vec{\mathrm{v}}) = \frac{1}{3}(\vec{\nabla}\cdot\vec{\mathrm{v}})\mathbf{1} + \dot{\boldsymbol{\varepsilon}}^d(\vec{\mathrm{v}})$$

where $\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathrm{v}})$ is the deviatoric strain rate tensor and (in Cartesian coordinates)

$$\vec{\nabla}\cdot\vec{\mathrm{v}} = \mathrm{div}(\vec{\mathrm{v}}) = \mathrm{tr}(\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}})) = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \tag{3.40}$$

where tr is the trace operator, we arrive at

$$\boldsymbol{\sigma} = (-p + \lambda(\vec{\nabla}\cdot\vec{\mathrm{v}}))\mathbf{1} + 2\eta\left[\frac{1}{3}(\vec{\nabla}\cdot\vec{\mathrm{v}})\mathbf{1} + \dot{\boldsymbol{\varepsilon}}^d\right] \tag{3.41}$$

$$= \left[-p + \left(\lambda + \frac{2}{3}\eta\right)(\vec{\nabla}\cdot\vec{\mathrm{v}})\right]\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}^d \tag{3.42}$$

Introducing the second viscosity $\zeta = \lambda + \frac{2}{3}\eta$:

$$\boldsymbol{\sigma} = \left[-p + \zeta(\vec{\nabla}\cdot\vec{\mathrm{v}})\right]\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}^d \tag{3.43}$$

$$= -\bar{p}\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}^d \tag{3.44}$$

---

[6]Note that often the dot is omitted and for example the ASPECT manual uses the ε notation.

[7]also sometimes called shear viscosity

The effect of the volume viscosity ζ is that the mechanical pressure $\bar{p}$ is not equivalent to the thermodynamic pressure p

$$\bar{p} = p - \zeta(\vec{\nabla} \cdot \vec{v}) \tag{3.45}$$

In other words: the isotropic average of the total stress is *not* the pressure term! This difference is usually neglected (and it is safe to do so, see [79, section 7.02.3.2.2]) by explicitly assuming $\zeta = 0$ (also called the Stokes assumption [1216, p256]), so that one can then refer to pressure as a single well-defined value. Note that in the case of an incompressible Newtonian Fluid, the strain rate tensor is deviatoric ($\text{tr}(\dot{\boldsymbol{\varepsilon}}(\vec{v})) = \text{div}(\vec{v}) = 0$) and the above considerations vanish.

Finally, for both compressible and incompressible flow, the stress tensor becomes simply

$$\boldsymbol{\sigma} = -p\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) = -p\mathbf{1} + \boldsymbol{\tau} \tag{3.46}$$

where $\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v})$ is the deviatoric stress tensor.

Remark. *On page 256 of Schubert, Turcotte and Olson [1216], equation 6.5.3, the authors write $\tau_{ii}/3 = k_B e_{ii}$ while stating that τ is deviatoric in equation 6.4.2. This is an obvious conflict of notations.*

3.6 The heat transport equation - energy conservation equation

3.7 The momentum conservation equations

As explained in Section 3.11, in Earth science applications the Navier-Stokes equations reduce to the Stokes equation:

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \rho\vec{g} = \vec{0} \tag{3.47}$$

Since

$$\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau} \tag{3.48}$$

it also writes

$$-\vec{\nabla}p + \vec{\nabla} \cdot \boldsymbol{\tau} + \rho\vec{g} = \vec{0} \tag{3.49}$$

Using the relationship $\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v})$ we arrive at

$$-\vec{\nabla}p + \vec{\nabla} \cdot (2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v})) + \rho\vec{g} = \vec{0} \tag{3.50}$$

The divergence of a tensor field in cylindrical coordinates (r, θ, z) has been obtained in Section 3.3.3. The equations of motion (3.47) becomes[8]

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{r\theta}}{\partial \theta} + \frac{\partial \sigma_{rz}}{\partial z} + \frac{1}{r}(\sigma_{rr} - \sigma_{\theta\theta}) + \rho g_r = 0 \tag{3.51}$$

$$\frac{\partial \sigma_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{\theta\theta}}{\partial \theta} + \frac{\partial \sigma_{\theta z}}{\partial z} + \frac{2}{r}\sigma_{r\theta} + \rho g_\theta = 0 \tag{3.52}$$

$$\frac{\partial \sigma_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{\theta z}}{\partial \theta} + \frac{\partial \sigma_{zz}}{\partial z} + \frac{1}{r}\sigma_{rz} + \rho g_z = 0 \tag{3.53}$$

---

[8]https://en.wikipedia.org/wiki/Linear_elasticity

3.8 The mass conservation equations

The mass conservation equation (often called continuity equation) is given by

$$\frac{D\rho}{Dt} + \rho\vec{\nabla}\cdot\vec{\mathbf{v}} = 0$$

or, since

$$\frac{D\rho}{Dt} = \frac{\partial\rho}{\partial t} + \vec{\mathbf{v}}\cdot\vec{\nabla}\rho$$

then

$$\frac{D\rho}{Dt} + \rho\vec{\nabla}\cdot\vec{\mathbf{v}} = \frac{\partial\rho}{\partial t} + \vec{\mathbf{v}}\cdot\vec{\nabla}\rho + \rho\vec{\nabla}\cdot\vec{\mathbf{v}} = 0$$

and finally:

$$\frac{\partial\rho}{\partial t} + \vec{\nabla}\cdot(\rho\vec{\mathbf{v}}) = 0 \tag{3.54}$$

In the case of an incompressible flow, then $\partial\rho/\partial t = 0$ and $\vec{\nabla}\rho = 0$, i.e. $D\rho/Dt = 0$ and the remaining equation is simply:

$$\vec{\nabla}\cdot\vec{\mathbf{v}} = 0$$

A vector field that is divergence-free is also called solenoidal[9].

In cylindrical coordinates (r, θ, ϕ) the continuity equation for an incompressible fluid is :

$$\frac{1}{r}\frac{\partial}{\partial r}(r\mathbf{v}_r) + \frac{1}{r}\frac{\partial\mathbf{v}_\theta}{\partial\theta} + \frac{\partial\mathbf{v}_z}{\partial z} = 0$$

In spherical coordinates (r, θ, ϕ) the continuity equation for an incompressible fluid is :

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\mathbf{v}_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\mathbf{v}_\theta\sin\theta) + \frac{1}{r\sin\theta}\frac{\partial\mathbf{v}_\phi}{\partial\phi} = 0 \tag{3.55}$$

3.9 The equations in Aspect manual

The following is lifted off the ASPECT manual. We focus on the system of equations in a $d = 2$- or $d = 3$-dimensional domain Ω that describes the motion of a highly viscous fluid driven by differences in the gravitational force due to a density that depends on the temperature. In the following, we largely follow the exposition of this material in Schubert, Turcotte and Olson [1216].

---

[9]https://en.wikipedia.org/wiki/Solenoidal_vector_field

Specifically, we consider the following set of equations for velocity $\vec{v}$, pressure p and temperature T:

$$-\vec{\nabla} \cdot \left[2\eta \left(\dot{\boldsymbol{\varepsilon}}(\vec{v}) - \frac{1}{3}(\vec{\nabla} \cdot \vec{v})\mathbf{1} \right) \right] + \vec{\nabla}p = \rho\vec{g} \qquad \text{in } \Omega, \tag{3.56}$$

$$\vec{\nabla} \cdot (\rho\vec{v}) = 0 \qquad \text{in } \Omega, \tag{3.57}$$

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla}T \right) - \vec{\nabla} \cdot k\vec{\nabla}T = \rho H$$

$$+ 2\eta \left(\dot{\varepsilon}(\vec{v}) - \frac{1}{3}(\vec{\nabla} \cdot \vec{v})\mathbf{1} \right) : \left(\dot{\varepsilon}(\vec{v}) - \frac{1}{3}(\vec{\nabla} \cdot \vec{v})\mathbf{1} \right) \tag{3.58}$$

$$+ \alpha T \left(\vec{v} \cdot \vec{\nabla}p \right) \qquad \text{in } \Omega,$$

where $\dot{\boldsymbol{\varepsilon}}(\vec{v}) = \frac{1}{2}(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T)$ is the symmetric gradient of the velocity (often called the *strain rate tensor*).

In this set of equations, (3.56) and (3.57) represent the compressible Stokes equations in which $\vec{v} = \vec{v}(\mathbf{x}, t)$ is the velocity field and $p = p(\mathbf{x}, t)$ the pressure field. Both fields depend on space $\mathbf{x}$ and time t. Fluid flow is driven by the gravity force that acts on the fluid and that is proportional to both the density of the fluid and the strength of the gravitational pull.

Coupled to this Stokes system is equation (3.58) for the temperature field $T = T(\mathbf{x}, t)$ that contains heat conduction terms as well as advection with the flow velocity $\mathbf{v}$. The right hand side terms of this equation correspond to

- internal heat production for example due to radioactive decay;

- friction (shear) heating;

- adiabatic compression of material;

In order to arrive at the set of equations in the ASPECT manual we need to

- neglect the $\partial p/\partial t$ _____ wrong rephrase

- neglect the $\partial \rho/\partial t$ in (3.54).

from equations above. A partial answer is given in the next section.

Also, their definition of the shear heating term Φ is:

$$\Phi = k_B(\vec{\nabla} \cdot \vec{v})^2 + 2\eta\dot{\varepsilon}^d : \dot{\varepsilon}^d$$

For many fluids the bulk viscosity k_B is very small and is often taken to be zero, an assumption known as the Stokes assumption: $k_B = \lambda + 2\eta/3 = 0$. Note that η is the dynamic viscosity and λ the second viscosity. Also,

$$\boldsymbol{\tau} = 2\eta\dot{\varepsilon} + \lambda(\boldsymbol{\nabla} \cdot \vec{v})\mathbf{1}$$

but since $k_B = \lambda + 2\eta/3 = 0$, then $\lambda = -2\eta/3$ so

$$\boldsymbol{\tau} = 2\eta\dot{\varepsilon} - \frac{2}{3}\eta(\boldsymbol{\nabla} \cdot \vec{v})\mathbf{1} = 2\eta\dot{\varepsilon}^d$$

3.10 Equations for thermal convection in an anelastic, compressible, self-gravitating spherical mantle

What follows is borrowed from Section 2.1 of Glišović *et al.* (2012) [497]. We start from the conservation mass, momentum and energy equations (the full Navier-Stokes equations):

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = \vec{0} \tag{3.59}$$

$$\rho \frac{D\vec{v}}{Dt} = \vec{\nabla} \cdot \boldsymbol{\sigma} + \rho \vec{g} \tag{3.60}$$

$$\rho C_p \frac{DT}{Dt} = \vec{\nabla} \cdot k \vec{\nabla} T + \alpha T \frac{Dp}{Dt} + \Phi + Q \tag{3.61}$$

In solving for the mantle flow field that satisfies the equation of momentum conservation, we incorporate all effects arising from self-gravitation and we must therefore explicitly consider the 3-D variation of gravity throughout Earth's interior. The gravitational acceleration is written as

$$\vec{g} = \vec{\nabla}\phi$$

where ϕ is Earth's gravitational potential which satisfies Poisson's equation

$$\Delta\phi = -4\pi\mathcal{G}\rho$$

The gravitational potential is expressed as

$$\phi = \phi_0(r) + \phi_1(r, \theta, \phi)$$

where the subscript 0 denotes a hydrostatic reference state, in which the structure of the mantle (density, gravity, pressure, temperature) varies with radius alone and the subscript 1 denotes all 3D perturbations arising from the thermal convection process. This decomposition makes sense in the context of a perfect sphere.

The total perturbed density and pressure fields in the mantle may similarly be expressed as

$$\rho = \rho_0(r) + \rho_1(r, \theta, \phi)$$

$$p = p_0(r) + p_1(r, \theta, \phi)$$

The equation of state relates the density perturbations to the temperature and pressure perturbations as follows

$$\rho_1 = \rho_0[1 - \alpha(T - T_0(r)) + K_T^{-1}(p - p_0(r))]$$

where K_T is the bulk modulus and the term $T_0(r)$ represents the horizontally averaged temperature (i.e. the geotherm) which varies with radius only. The effects of compressibility on the density are found to be at least two orders of magnitude smaller than the effects of temperature variations. Therefore, the last term of this equation is often neglected. Note that this expression is a first order expansion of any Equation of State.

Also, this equation can be misleading if one forgets that the parameters α and K_T cannot be constant but must be related through Maxwell relations (for example, their definitions

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P = -\frac{1}{\rho}\left(\frac{\partial \rho}{\partial T}\right)_P$$

and

$$K_T = -V \left(\frac{\partial P}{\partial V} \right)_T = \rho \left(\frac{\partial P}{\partial \rho} \right)_T$$

imply that

$$\frac{\partial(\alpha\rho)}{\partial P} = -\frac{\partial(\rho/K_T)}{\partial T}$$

Some models can be found in the geophysical literature in which assumptions made inconsistently about thermodynamic parameters (either constant or depth-dependent) violate the Maxwell rules.

Important simplifications are made assuming the anelastic-liquid approximation (e.g. Jarvis & McKenzie (1980) [672], Solheim & Peltier (1990) [1258]). This approximation is justified because the velocities associated with mantle convection are very slow compared to the local sound speed and hence acoustic waves cannot be generated by the slow changes in the mantle pressure field. We therefore neglect the time derivative of density, thereby eliminating sound waves:

$$\frac{\partial \rho}{\partial t} \simeq 0$$

For the same reason, the pressure distribution may be considered (to first-order accuracy) as the pressure of a fluid in hydrostatic equilibrium which yields

$$\frac{Dp}{Dt} = \frac{\partial p}{\partial t} + \vec{v} \cdot \vec{\nabla} p \simeq -u_r \rho_0(r) g_0(r)$$

The equations are then rewritten in terms of dimensionless variables according to the relations:

$$r' = \frac{r}{d} \tag{3.62}$$

$$\vec{v}' = \frac{\vec{v}}{U} \tag{3.63}$$

$$t' = \frac{U}{d/t} \tag{3.64}$$

$$T' = \frac{T}{\Delta T} \tag{3.65}$$

$$\rho' = \frac{\rho}{\rho_{0s}} \tag{3.66}$$

$$g' = \frac{g}{g_{0s}} \tag{3.67}$$

$$\phi' = \frac{\phi}{g_{0s}d} \tag{3.68}$$

$$\alpha' = \frac{\alpha}{\alpha_s} \tag{3.69}$$

$$p' = \frac{p}{\alpha_s \Delta T \rho_{0s} g_{0s} d} \tag{3.70}$$

$$\tau_{ij} = \frac{\tau_{ij}}{\alpha_s \Delta T \rho_{0s} g_{0s} d} \tag{3.71}$$

$$\eta' = \frac{\eta}{\eta_s} \tag{3.72}$$

$$k' = \frac{k}{k_s} \tag{3.73}$$

$$Q' = \frac{Qd^2}{k_s \Delta T} \tag{3.74}$$

$$U = \frac{\rho_{0s} g_{0s} \alpha_s \Delta T d^2}{\eta_s} \tag{3.75}$$

in which the primes represent the dimensionless variables, the subscript s means that we consider the surface value of the variable to which it is applied. The length scale d and temperature scale ΔT are respectively the depth of the mantle and the difference of temperature between the bottom and the top of the mantle.

Often one deals with dimensionless variables and the primes are dropped for notational convenience (this is the case in what follows).

It is a tedious but trivial exercise to show that the dimensionless equation of conservation of momentum is then written as follows:

$$\rho \frac{\mathsf{Ra}_s}{Pr_s} \frac{D\vec{\mathsf{v}}}{Dt} = \frac{\rho}{\alpha_s \Delta T} \vec{\nabla}\phi - \vec{\nabla}p + \vec{\nabla} \cdot \boldsymbol{\tau}$$

in which we introduce the surface Rayleigh Ra_s and Prandtl Pr_s numbers defined, respectively, by

$$\mathsf{Ra}_s = \frac{\rho_{0s}^2 C_p g_{0s} \alpha_s \Delta T d^3}{k_s \eta_s} \qquad Pr_s = \frac{\eta_s C_p}{k_s}$$

Because of the very high viscosity of mantle rocks, the left-hand term is smaller than the other terms by several orders of magnitude and may therefore be neglected. This important simplification is called the infinite Prandtl number approximation.

The equation of energy conservation may also be rewritten in terms of the surface Rayleigh number, as follows

$$\frac{DT}{Dt} = \frac{1}{\rho \mathsf{Ra}_s} \left(\vec{\nabla} \cdot k\vec{\nabla}T + Q \right) + \frac{Di}{\rho} \left(-\alpha T \frac{Dp}{Dt} + \Phi \right)$$

where Di is the dissipation number (see Peltier (1972) [1047]) which measures the importance of compression work and frictional heating, and it is defined as

$$Di = \frac{\alpha_s g_{0s} d}{C_p}$$

Di also measures the ratio of the depth of mantle convection (d) to the adiabatic scale height ($C_p/\alpha g_0$) and for whole-mantle convection is close to order 1 (see Jarvis & McKenzie (1980) [672]).

After simplifications, the dimensionless system of governing equations is written as

$$\vec{\nabla} \cdot (\rho_0 \vec{\mathsf{v}}) = 0 \tag{3.76}$$

$$\frac{\rho}{\alpha_s \Delta T} \vec{\nabla}\phi - \vec{\nabla}p + \vec{\nabla} \cdot \boldsymbol{\tau} = \vec{0} \tag{3.77}$$

$$\frac{\partial T}{\partial t} + \vec{\mathsf{v}} \cdot \vec{\nabla}T = \frac{1}{\rho_0 \mathsf{Ra}_s} \left(\vec{\nabla} \cdot k\vec{\nabla}T + Q \right) + \frac{Di}{\rho_0} (-\alpha T \rho_0 g_0 u_r + \Phi) \tag{3.78}$$

with

$$\Delta\phi = -4\pi\mathcal{G}\rho$$

$$\rho_1 = \rho_0 (1 - \alpha(T - T_0(r)))$$

VERIFY all this !

3.11 Non-dimensionalisation of the Navier-Stokes equations

3.11.1 Approach # 1 - isothermal flow

We define (see for instance Massimi et $al.$ (2006) [888]) four reference quantities which are relevant for geodynamics[10]:

- a reference viscosity value $\underline{\eta} = 10^{20}$ Pa s

- a reference mass density $\underline{\rho} = 1000$ kg m^{-3}

- a reference time $\underline{t} = 1$Myr $\simeq 3.15 \cdot 10^{13}$ s

- a reference length $\underline{l} = 1000$ m

- a reference gravity $\underline{g} = 9.81$ m s^{-2}

It follows that a reference pressure can be obtained:

$$\underline{p} = \underline{\rho}\underline{g}\underline{l} = 9.81 \cdot 10^6 \text{ Pa}$$

Note that there is unfortunately no natural selection for the pressure scale. We could also have used $\underline{p} = \underline{\rho}\underline{v}^2$ where dynamic effects are dominant i.e. high velocity flows, or $\underline{p} = \underline{\eta}\underline{v}/\underline{l}$ where viscous effects are dominant i.e. creeping flows (which is the case in geodynamics). The definition of a reference velocity is more straightforward:

$$\underline{v} = \frac{\underline{l}}{\underline{t}} = 1 \text{ mm yr}^{-1}$$

We define dimensionless variables through:

$$x = \frac{x}{\underline{l}} \qquad y = \frac{y}{\underline{l}} \qquad z = \frac{z}{\underline{l}} \qquad \vec{v}' = \frac{\vec{v}}{\underline{v}} \qquad t = \frac{t}{\underline{t}} \qquad \eta = \frac{\eta}{\underline{\eta}} \qquad g = \frac{g}{\underline{g}}$$

where the teal color indicates dimensionless values.

Consequently, time and space derivatives will be rescaled as follows:

$$\vec{\nabla} = \underline{l}\,\vec{\nabla} \qquad \partial_t = \underline{t}\,\partial_t$$

Using this scaling relations the Navier-Stokes equation become:

$$\frac{\underline{\rho}\underline{l}}{\underline{t}^2}\frac{\partial\vec{v}}{\partial t} + \frac{\underline{\rho}\underline{l}}{\underline{t}^2}(\vec{v}\cdot\vec{\nabla})\vec{v} = -\underline{\rho}\underline{g}\vec{\nabla}p + \frac{\underline{\eta}}{\underline{l}\underline{t}}\vec{\nabla}\cdot\eta(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T) + \rho\vec{g}$$

I make $\rho = \rho/\underline{\rho}$ appear in the left hand side:

$$\frac{\underline{\rho}\underline{l}}{\underline{t}^2}\rho\frac{\partial\vec{v}}{\partial t} + \frac{\underline{\rho}\underline{l}}{\underline{t}^2}\rho(\vec{v}\cdot\vec{\nabla})\vec{v} = -\underline{\rho}\underline{g}\vec{\nabla}p + \frac{\underline{\eta}}{\underline{l}\underline{t}}\vec{\nabla}\cdot\eta(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T) + \rho\vec{g}$$

which we can divide by $\underline{\rho}\underline{l}/\underline{t}^2$ to obtain:

$$\rho\left(\frac{\partial\vec{v}}{\partial t} + (\vec{v}\cdot\vec{\nabla})\vec{v}\right) = -\frac{\underline{g}\underline{t}^2}{\underline{l}}\vec{\nabla}p + \frac{\underline{\eta}\underline{t}}{\underline{\rho}\underline{l}^2}\vec{\nabla}\cdot\eta(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T) + \frac{\underline{t}^2}{\underline{l}}\rho\vec{g}$$

One can recognise in this equation the Reynolds and Froude non-dimensional numbers (the ratio between the inertial and viscous forces, and the ratio between buoyancy and inertial forces respectively).

$$Re = \frac{\underline{\rho}\underline{l}^2}{\underline{\eta}\underline{t}} \qquad Fr = \frac{\underline{l}}{\underline{g}\underline{t}^2}$$

---

[10]Note that in the paper the authors conflate ρ and $\tilde{\rho}$ which prevents them from non-dimensionalising all terms as we do here.

From this we conclude that inertial forces in the Earth's mantle are small compared to viscous forces. We can then write:

$$\boxed{\rho\left(\frac{\partial\vec{v}}{\partial t}+(\vec{v}\cdot\vec{\nabla})\vec{v}\right)=-\frac{1}{Fr}\vec{\nabla}p+\frac{1}{Re}\vec{\nabla}\cdot\eta(\vec{\nabla}\vec{v}+\vec{\nabla}\vec{v}^T)+\frac{1}{Fr}\vec{g}}$$

In our case, given the definitions taken above, we have:

$$Re\simeq 3.174\cdot 10^{-24}\qquad Fr\simeq 1.027\cdot 10^{-25}$$

so that the inertial terms can be dropped from the momentum equation (thereby yielding the dimensionless Stokes equations):

$$\vec{\nabla}\cdot\eta(\vec{\nabla}\vec{v}+\vec{\nabla}\vec{v}^T)-\frac{Re}{Fr}\vec{\nabla}p+\frac{Re}{Fr}\vec{g}=0$$

Note that in our case $Re/Fr\simeq 30.5$.

3.11.2   Approach # 2 - Temperature dependent

Let us now consider a box heated from below and cooled from above. We define 4 fundamental reference quantities:

- a length L_{ref} (m), (L)

- a temperature T_{ref} (K), (θ)

- a viscosity η_{ref} (Pa s), ($ML^{-1}T^{-1}$)

- a thermal diffusion coefficient κ_{ref} (m^2 s^{-1}), (L^2T^{-1})

From these reference quantities one can form secondary ones, such as

- a time $t_{ref}=L_{ref}^2/\kappa_{ref}$ (aka the diffusion time)

- a velocity $v_{ref}=L_{ref}/t_{ref}=\kappa_{ref}/L_{ref}$

- an acceleration $g_{ref}=v_{ref}/t_{ref}=\kappa_{ref}^2/L_{ref}^3$

- a strain rate $\dot{\varepsilon}_{ref}=t_{ref}^{-1}=\kappa_{ref}/L_{ref}^2$

- a pressure $p_{ref}=\eta_{ref}\dot{\varepsilon}_{ref}=\eta_{ref}t_{ref}^{-1}$

- a reference density $\rho_{ref}=\eta_{ref}L_{ref}t_{ref}/L_{ref}^3=\eta_{ref}L_{ref}^{-2}t_{ref}$

- a reference mass $M_{ref}=\eta_{ref}L_{ref}t_{ref}$

- a reference energy $E_{ref}=\eta_{ref}L_{ref}t_{ref}\frac{L_{ref}^2}{t_{ref}^2}=\eta_{ref}\frac{L_{ref}^3}{t_{ref}}$

- a reference heat conductivity[11] $k_{ref}=E_{ref}/t_{ref}/L_{ref}/T_{ref}=\eta_{ref}L_{ref}^2/t_{ref}^2/T_{ref}$

- a reference heat capacity[12] $C_{ref}=E_{ref}/M_{ref}/T_{ref}=L_{ref}^2/t_{ref}^2/T_{ref}$

- a reference heat production coefficient[13] $H_{ref}=E_{ref}/t_{ref}/M_{ref}=\frac{L_{ref}^2}{t_{ref}^3}$

- a reference heat flux[14] $q_{ref}=\eta_{ref}L_{ref}t_{ref}^{-2}$

---

[11]Units: W/m/K
[12]Units: J/kg/K
[13]Units: W/kg
[14]Units: W m^{-2}, or kg s^{-3}

We define dimensionless quantities as follows:

$$x = \frac{x}{L_{ref}} \quad \vec{\mathbf{v}} = \frac{\vec{\mathbf{v}}}{\mathbf{v}_{ref}} \quad t = \frac{t}{t_{ref}} \quad \eta = \frac{\eta}{\eta_{ref}} \quad g = \frac{g}{g_{ref}} \quad k = \frac{k}{k_{ref}} \quad C_p = \frac{C_p}{C_{ref}} \quad \rho = \frac{\rho}{\rho_{\text{ref}}} \qquad (3.79)$$

$$H = \frac{H}{H_{ref}} \quad \vec{\nabla} = L_{ref}\,\vec{\nabla} \quad \partial_t = t_{ref}\,\partial_t \quad T = \frac{T}{T_{ref}} \quad \dot{\varepsilon} = \dot{\varepsilon}\,t_{ref} \qquad (3.80)$$

We start from the standard Navier-Stokes equation[15]

$$\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot(2\eta\dot{\varepsilon}) + \rho\vec{g}$$

and assume that the density is temperature-dependent (Boussinesq approximation) so that

$$\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot(2\eta\dot{\varepsilon}) + \rho_0(1 - \alpha T)\vec{g}$$

and remove the hydrostatic pressure (although we keep using p for simplicity, p is now the dynamic pressure):

$$\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot(2\eta\dot{\varepsilon}) - \rho_0\alpha T\vec{g}$$

We divide this equation by $p_{ref} = \eta_{ref}\dot{\varepsilon}_{ref}$:

$$\frac{1}{\eta_{ref}\dot{\varepsilon}_{ref}}\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\frac{\eta}{\eta_{ref}}\frac{\dot{\varepsilon}}{\dot{\varepsilon}_{ref}} - \frac{\rho_0\alpha T\vec{g}}{\eta_{ref}\dot{\varepsilon}_{ref}}$$

Let us call $\vec{e}$ the positive vertical vector ($\vec{e}_z$ in Cartesian coordinates, $\vec{e}_r$ in spherical coordinates), then $\vec{g} = -g_0\vec{e}$ and we can write (using $\dot{\varepsilon}_{ref} = \kappa_{ref}/L_{ref}^2$)

$$\frac{1}{\eta_{ref}\dot{\varepsilon}_{ref}}(\rho_{ref}\rho)\frac{D(\mathbf{v}_{ref}\vec{\mathbf{v}})}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \frac{\rho_0\alpha T g_0}{\eta_{ref}(\kappa_{ref}/L_{ref}^2)}\vec{e}$$

Finally, dividing by L_{ref}^{-1} (i.e. multiplying by L_{ref}) yields

$$\frac{\mathbf{v}_{ref}\rho_{ref}L_{ref}}{\eta_{ref}\dot{\varepsilon}_{ref}}\rho\frac{D\vec{\mathbf{v}}}{t_{ref}Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \frac{\rho_0\alpha(TT_{ref})g_0L_{ref}^3}{\eta_{ref}\kappa_{ref}}\vec{e}$$

and finally (using $\mathbf{v}_{ref} = L_{ref}/t_{ref}$)

$$\frac{\rho_{ref}\kappa_{ref}}{\eta_{ref}}\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \frac{\rho_0\alpha T_{ref}g_0L_{ref}^3}{\eta_{ref}\kappa_{ref}}T\vec{e}$$

In the context of a system with a temperature difference ΔT between the bottom and top boundaries separated by a distance H, one would then take $T_{ref} = \Delta T$ and $L_{ref} = H$ so that the equation becomes:

$$\underbrace{\frac{\rho_{ref}\kappa_{ref}}{\eta_{ref}}}_{\mathsf{Pr}^{-1}}\rho\frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \underbrace{\frac{\rho_0\alpha\Delta T g_0 H^3}{\eta_{ref}\kappa_{ref}}}_{\mathsf{Ra}}T\vec{e}$$

and we obviously recover the classical definition of the Rayleigh number.

---

[15]https://en.wikipedia.org/wiki/Navier-Stokes_equations

$$\frac{1}{\mathsf{Pr}}\rho\frac{D\vec{v}}{Dt} = -\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \mathsf{Ra}T\vec{e}$$

On the left side of the equation we recognize the (inverse of the) Prandlt number $\mathsf{Pr} = \frac{\eta}{\rho\kappa}$. We can estimate the dimensionless number before the inertial term for Earth geodynamics:

$$\mathsf{Pr} \simeq \frac{10^{20-23}}{3000\cdot 10^{-6}} >> 10^{23}$$

Its inverse is then extremely small and this is why we neglect the inertial terms in mantle modelling.

Note that if the fluid is isoviscous, one can then set $\eta_{ref} = \eta = \eta_0$ and then $\eta = 1$

Turning now to the continuity equation $\vec{\nabla}\cdot\vec{v} = 0$, it is trivial to show that $\vec{\nabla}\cdot\vec{v} = 0$. Finally, starting from the simple heat transport equation:

$$\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T = \kappa\Delta T$$

We divide each side by T_{ref} so that

$$\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T = \kappa\Delta T$$

We now divide each side by the reference velocity v_{ref} and we obtain

$$\frac{L_{ref}}{\kappa_{ref}}\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T = \frac{L_{ref}}{\kappa_{ref}}\kappa\Delta T$$

We multiply each side by L_{ref} and we finally get

$$\frac{L_{ref}^2}{\kappa_{ref}}\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T = \kappa\Delta T$$

and finally

$$\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T = \kappa\Delta T$$

The set of dimensionless equations is then:

$$
\begin{aligned}
-\vec{\nabla}p + \vec{\nabla}\cdot 2\eta\dot{\varepsilon} + \mathsf{Ra}T\vec{e} &= \vec{0} & (3.81)\\
\vec{\nabla}\cdot\vec{v} &= 0 & (3.82)\\
\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T &= \kappa\Delta T & (3.83)
\end{aligned}
$$

Looking now at the Extended Boussinesq Approximation (EBA), we have to conside two additional terms in the energy equation:

- the shear heating Φ (See Eq.(18.84)) $\Phi = 2\eta\dot{\varepsilon}^d : \dot{\varepsilon}^d$

- the adiabatic heating $\alpha T\vec{v}\cdot\vec{\nabla}p$

We start this time from

$$\rho C_p\left(\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T\right) - \vec{\nabla}\cdot k\vec{\nabla}T = \rho H + 2\eta\dot{\varepsilon}^d : \dot{\varepsilon}^d + \alpha T\vec{v}\cdot\vec{\nabla}p$$

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \vec{\nabla} \cdot k\vec{\nabla} T = \rho H + 2\eta \dot{\varepsilon}^d : \dot{\varepsilon}^d + \alpha T \vec{v} \cdot \vec{\nabla} p$$

$$\rho C_p \frac{T_{ref}}{t_{ref}} \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \frac{T_{ref}}{L_{ref}^2} \vec{\nabla} \cdot k\vec{\nabla} T = \rho_{ref} \rho H + \frac{\eta_{ref}}{t_{ref}^2} 2\eta \dot{\varepsilon}^d : \dot{\varepsilon}^d + \frac{p_{ref}}{t_{ref}} \alpha T \vec{v} \cdot \vec{\nabla} p$$

we then use $p_{ref} = \eta_{ref} t_{ref}^{-1}$ and $\rho_{ref} = \eta_{ref} L_{ref}^{-2} t_{ref}$

$$\rho_{ref} C_{ref} \frac{T_{ref}}{t_{ref}} \rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \frac{T_{ref} k_{ref}}{L_{ref}^2} \vec{\nabla} \cdot k\vec{\nabla} T = \frac{\eta_{ref}}{L_{ref}^2} t_{ref} \frac{L_{ref}^2}{t_{ref}^3} \rho H + \frac{\eta_{ref}}{t_{ref}^2} 2\eta \dot{\varepsilon}^d : \dot{\varepsilon}^d + \frac{\eta_{ref}}{t_{ref}^2} \alpha T \vec{v} \cdot \vec{\nabla} p$$

or, multiplying all by t_{ref}^2 / η_{ref}:

$$\frac{t_{ref}^2}{\eta_{ref}} \rho_{ref} C_{ref} \frac{T_{ref}}{t_{ref}} \rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \frac{t_{ref}^2}{\eta_{ref}} \frac{T_{ref} k_{ref}}{L_{ref}^2} \vec{\nabla} \cdot k\vec{\nabla} T = \rho H + 2\eta \dot{\varepsilon}^d : \dot{\varepsilon}^d + \alpha T \vec{v} \cdot \vec{\nabla} p$$

We then make use of $C_{ref} = L_{ref}^2 / t_{ref}^2 / T_{ref}$ and $k_{ref} = \eta_{ref} L_{ref}^2 / t_{ref}^2 / T_{ref}$ to arrive at

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \vec{\nabla} \cdot k\vec{\nabla} T = \rho H + 2\eta \dot{\varepsilon}^d : \dot{\varepsilon}^d + \alpha T \vec{v} \cdot \vec{\nabla} p$$

3.12 The Navier-Stokes equations in cylindrical coordinates

In cylindrical coordinates, (r, θ, z), the continuity equation for an incompressible fluid is

$$\frac{1}{r}\frac{\partial}{\partial r}(r\mathbf{v}_r) + \frac{1}{r}\frac{\partial}{\partial \theta}(\mathbf{v}_\theta) + \frac{\partial \mathbf{v}_z}{\partial z} = 0 \tag{3.84}$$

or

$$\frac{\partial \mathbf{v}_r}{\partial r} + \frac{\mathbf{v}_r}{r} + \frac{1}{r}\frac{\partial}{\partial \theta}(\mathbf{v}_\theta) + \frac{\partial \mathbf{v}_z}{\partial z} = 0 \tag{3.85}$$

The Navier-Stokes equations of motion for an incompressible fluid with uniform viscosity are:

$$\begin{aligned}
\rho\left(\frac{D\mathbf{v}_r}{Dt} - \frac{\mathbf{v}_\theta^2}{r}\right) &= -\frac{\partial p}{\partial r} + f_r + \eta\left(\Delta \mathbf{v}_r - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right) \\
\rho\left(\frac{D\mathbf{v}_\theta}{Dt} + \frac{\mathbf{v}_\theta \mathbf{v}_r}{r}\right) &= -\frac{1}{r}\frac{\partial p}{\partial \theta} + f_\theta + \eta\left(\Delta \mathbf{v}_\theta - \frac{\mathbf{v}_\theta}{r^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta}\right) \\
\rho\frac{D\mathbf{v}_z}{Dt} &= -\frac{\partial p}{\partial z} + f_z + \eta\Delta \mathbf{v}_z
\end{aligned} \tag{3.86}$$

where the Lagrangian or material derivative is

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v}_r\frac{\partial}{\partial r} + \frac{\mathbf{v}_\theta}{r}\frac{\partial}{\partial \theta} + \mathbf{v}_z\frac{\partial}{\partial z}$$

and the Laplacian operator is

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + \frac{\partial^2}{\partial z^2} \tag{3.87}$$

and for an incompressible, Newtonian fluid

$$\sigma_{rr} = -p + 2\eta\frac{\partial \mathbf{v}_r}{\partial r} \tag{3.88}$$

$$\sigma_{\theta\theta} = -p + 2\eta\left(\frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta} + \frac{u}{r}\right) \tag{3.89}$$

$$\sigma_{zz} = -p + 2\eta\frac{\partial \mathbf{v}_z}{\partial z} \tag{3.90}$$

$$\sigma_{rz} = \eta\left(\frac{\partial \mathbf{v}_r}{\partial z} + \frac{\partial \mathbf{v}_z}{\partial r}\right) \tag{3.91}$$

$$\sigma_{r\theta} = \eta\left(\frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial \theta} + \frac{\partial \mathbf{v}_\theta}{\partial r} - \frac{\mathbf{v}_\theta}{r}\right) \tag{3.92}$$

$$\sigma_{\theta z} = \eta\left(\frac{1}{r}\frac{\partial \mathbf{v}_z}{\partial \theta} + \frac{\partial \mathbf{v}_\theta}{\partial z}\right) \tag{3.93}$$

3.13 The Stokes equations in spherical coordinates

In spherical coordinates, (r, θ, ϕ), the continuity equation for an incompressible fluid is

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \mathbf{v}_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(\mathbf{v}_\theta \sin\theta) + \frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_\phi}{\partial \phi} = 0 \tag{3.94}$$

Concerning the momentum equation, we start from

$$\vec{\nabla}\cdot\boldsymbol{\sigma} + \vec{f} \;=\; \vec{0} \tag{3.95}$$

The buoyancy force $\vec{f}$ is nearly always given by $\vec{f} = \rho\vec{g} = -\rho g\, \vec{e}_r$ (with $g > 0$), i.e. $f_\phi = f_\theta = 0$, and then

$$-\vec{\nabla}p + \vec{\nabla}\cdot\boldsymbol{\tau} - \rho g\vec{e}_r = \vec{0}$$

or,

$$
\begin{aligned}
-(\vec{\nabla}p)_r + (\vec{\nabla}\cdot\boldsymbol{\tau})_r &= \rho g \\
-(\vec{\nabla}p)_\theta + (\vec{\nabla}\cdot\boldsymbol{\tau})_\theta &= 0 \\
-(\vec{\nabla}p)_\phi + (\vec{\nabla}\cdot\boldsymbol{\tau})_\phi &= 0
\end{aligned}
$$

The pressure gradient is simply given by:

$$
\begin{aligned}
(\vec{\nabla}p)_r &= \frac{\partial p}{\partial r} \\
(\vec{\nabla}p)_\theta &= \frac{1}{r}\frac{\partial p}{\partial \theta} \\
(\vec{\nabla}p)_\phi &= \frac{1}{r\sin\theta}\frac{\partial p}{\partial \phi}
\end{aligned}
$$

We now turn to the remaining three components of the divergence of deviatoric stress in spherical coordinates r, θ, ϕ, which are given by[16]

$$
\begin{aligned}
(\vec{\nabla}\cdot\boldsymbol{\tau})_r &= \frac{\partial \tau_{rr}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta r}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi r}}{\partial \phi} + \frac{2\tau_{rr} - \tau_{\theta\theta} - \tau_{\phi\phi}}{r} + \frac{\tau_{\theta r}\cot\theta}{r} \\
(\vec{\nabla}\cdot\boldsymbol{\tau})_\theta &= \frac{\partial \tau_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta\theta}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi\theta}}{\partial \phi} + \frac{3\tau_{\theta r} + (\tau_{\theta\theta} - \tau_{\phi\phi})\cot\theta}{r} \\
(\vec{\nabla}\cdot\boldsymbol{\tau})_\phi &= \frac{\partial \tau_{r\phi}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta\phi}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi\phi}}{\partial \phi} + \frac{3\tau_{r\phi} + 2\tau_{\phi\theta}\cot\theta}{r}
\end{aligned}
\tag{3.96}
$$

And finally the momentum equation writes:

$$-\frac{\partial p}{\partial r} + \frac{\partial \tau_{rr}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta r}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi r}}{\partial \phi} + \frac{2\tau_{rr} - \tau_{\theta\theta} - \tau_{\phi\phi}}{r} + \frac{\tau_{\theta r}\cot\theta}{r} \;=\; \rho g$$

$$-\frac{1}{r}\frac{\partial p}{\partial \theta} + \frac{\partial \tau_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta\theta}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi\theta}}{\partial \phi} + \frac{3\tau_{\theta r} + (\tau_{\theta\theta} - \tau_{\phi\phi})\cot\theta}{r} \;=\; 0$$

$$-\frac{1}{r\sin\theta}\frac{\partial p}{\partial \phi} + \frac{\partial \tau_{r\phi}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta\phi}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \tau_{\phi\phi}}{\partial \phi} + \frac{3\tau_{r\phi} + 2\tau_{\phi\theta}\cot\theta}{r} \;=\; 0 \tag{3.97}$$

---

[16]Would be nice to have a ref here

The deviatoric stress tensor components are

$$\tau_{rr} = 2\eta\left[\dot{\varepsilon}_{rr} - \frac{1}{3}(\dot{\varepsilon}_{rr} + \dot{\varepsilon}_{\theta\theta} + \dot{\varepsilon}_{\phi\phi})\right]$$

$$\tau_{\theta\theta} = 2\eta\left[\dot{\varepsilon}_{\theta\theta} - \frac{1}{3}(\dot{\varepsilon}_{rr} + \dot{\varepsilon}_{\theta\theta} + \dot{\varepsilon}_{\phi\phi})\right]$$

$$\tau_{\phi\phi} = 2\eta\left[\dot{\varepsilon}_{\phi\phi} - \frac{1}{3}(\dot{\varepsilon}_{rr} + \dot{\varepsilon}_{\theta\theta} + \dot{\varepsilon}_{\phi\phi})\right]$$

$$\tau_{r\theta} = 2\eta\dot{\varepsilon}_{r\theta} \tag{3.98}$$

$$\tau_{r\phi} = 2\eta\dot{\varepsilon}_{r\phi} \tag{3.99}$$

$$\tau_{\theta\phi} = 2\eta\dot{\varepsilon}_{\theta\phi} \tag{3.100}$$

with

$$\dot{\varepsilon}_{rr} = \frac{\partial \mathbf{v}_r}{\partial r}$$

$$\dot{\varepsilon}_{\theta\theta} = \frac{\mathbf{v}_r}{r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta}$$

$$\dot{\varepsilon}_{\phi\phi} = \frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_\phi}{\partial \phi} + \frac{\mathbf{v}_r}{r} + \frac{\mathbf{v}_\theta\cot\theta}{r}$$

$$\dot{\varepsilon}_{\theta r} = \dot{\varepsilon}_{r\theta} = \frac{1}{2}\left(r\frac{\partial}{\partial r}(\frac{\mathbf{v}_\theta}{r}) + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial \theta}\right)$$

$$\dot{\varepsilon}_{\phi r} = \dot{\varepsilon}_{r\phi} = \frac{1}{2}\left(\frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_r}{\partial \phi} + r\frac{\partial}{\partial r}(\frac{\mathbf{v}_\phi}{r})\right)$$

$$\dot{\varepsilon}_{\phi\theta} = \dot{\varepsilon}_{\theta\phi} = \frac{1}{2}\left(\frac{\sin\theta}{r}\frac{\partial}{\partial \theta}(\frac{\mathbf{v}_\phi}{\sin\theta}) + \frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_\theta}{\partial \phi}\right)$$

We go further by assuming the fluid to be incompressible (i.e. $\dot{\varepsilon}_{rr} + \dot{\varepsilon}_{\theta\theta} + \dot{\varepsilon}_{\phi\phi} = 0$) and then:

$$\tau_{rr} = 2\eta\dot{\varepsilon}_{rr} = 2\eta\frac{\partial \mathbf{v}_r}{\partial r}$$

$$\tau_{\theta\theta} = 2\eta\dot{\varepsilon}_{\theta\theta} = 2\eta\left(\frac{\mathbf{v}_r}{r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right)$$

$$\tau_{\phi\phi} = 2\eta\dot{\varepsilon}_{\phi\phi} = 2\eta\left(\frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_\phi}{\partial \phi} + \frac{\mathbf{v}_r}{r} + \frac{\mathbf{v}_\theta\cot\theta}{r}\right)$$

$$\tau_{r\theta} = 2\eta\dot{\varepsilon}_{r\theta} = \eta\left(r\frac{\partial}{\partial r}(\frac{\mathbf{v}_\theta}{r}) + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial \theta}\right)$$

$$\tau_{r\phi} = 2\eta\dot{\varepsilon}_{r\phi} = \eta\left(\frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_r}{\partial \phi} + r\frac{\partial}{\partial r}(\frac{\mathbf{v}_\phi}{r})\right)$$

$$\tau_{\theta\phi} = 2\eta\dot{\varepsilon}_{\theta\phi} = \eta\left(\frac{\sin\theta}{r}\frac{\partial}{\partial \theta}(\frac{\mathbf{v}_\phi}{\sin\theta}) + \frac{1}{r\sin\theta}\frac{\partial \mathbf{v}_\theta}{\partial \phi}\right)$$

Inserting these expressions in Eq. (3.97) is a cumbersome affair... Under the assumption that the

fluid is also isoviscous, we get[17]

$$
\begin{aligned}
\rho g &= -\frac{\partial p}{\partial r} + \eta\left(\Delta \mathbf{v}_r - \frac{2\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta} - \frac{2\mathbf{v}_\theta \cot\theta}{r^2} - \frac{2}{r^2 \sin\theta}\frac{\partial \mathbf{v}_\phi}{\partial \phi}\right) \\
0 &= -\frac{1}{r}\frac{\partial p}{\partial \theta} + \eta\left(\Delta \mathbf{v}_\theta + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2 \sin^2\theta} - \frac{2\cot\theta}{r^2 \sin\theta}\frac{\partial \mathbf{v}_\phi}{\partial \phi}\right) \\
0 &= -\frac{1}{r\sin\theta}\frac{\partial p}{\partial \phi} + \eta\left(\Delta \mathbf{v}_\phi + \frac{2}{r^2 \sin\theta}\frac{\partial \mathbf{v}_r}{\partial \phi} - \frac{\mathbf{v}_\phi}{r^2 \sin^2\theta} + \frac{2\cot\theta}{r^2 \sin\theta}\frac{\partial \mathbf{v}_\theta}{\partial \phi}\right)
\end{aligned}
$$

(3.101)

and the Laplacian operator is

$$
\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \phi^2}
$$

> Before being used these equations should be checked against multiple sources.

3.14 The equations for axisymmetric geometries

In what follows we are concerned with incompressible flow. In some cases the assumption can be made that the object we wish to sudy has an axisymmetric geometry, for example a plume:
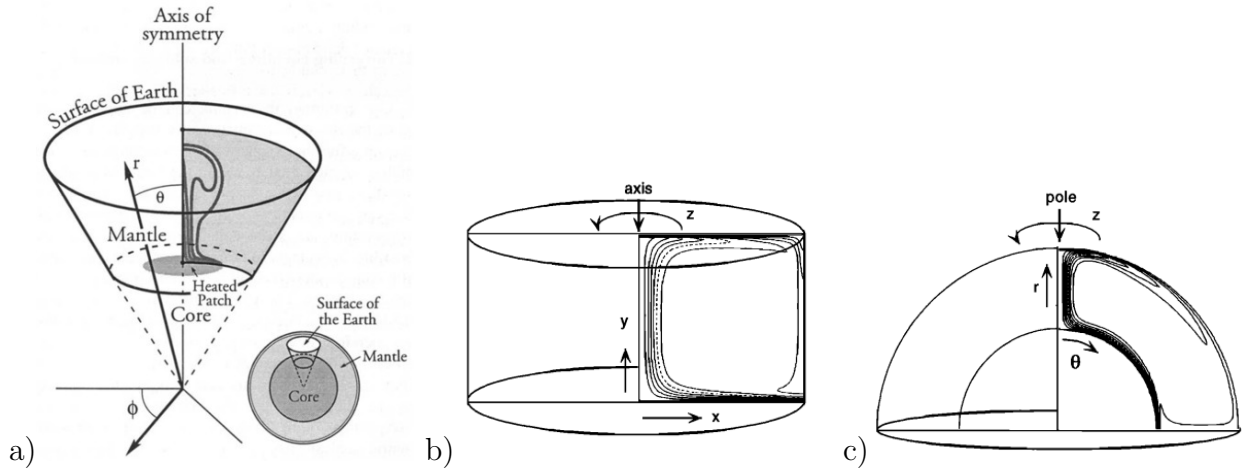
a)Taken from Kellogg and King [730] (1997); b,c) Taken from Leitch, Steinbach, and Yuen [808] (1996).

Looking at the figure above we see that there are in fact two cases: axisymmetry in cylindrical coordinates (b) and axisymmetry in spherical coordinates (c).

As mentioned in Kellogg and King [730] (1997): "By imposing axisymmetry, we restrict the problem to two degrees of freedom, reducing the computational effort significantly over 3D calculations." However, Redmond and King [1119] (2004) also mention: "An important caveat of axisymmetric calculations is that there are no variations in the ϕ direction (i.e., there are no ϕ derivatives in the governing equations). Thus, as we get further from the pole, the results become increasingly less physical. Downwelling drips off the pole are actually downwelling doughnuts that follow the entire small circle. In a fully 3D calculation, this doughnut feature would in reality be a drip."

See Section 10.8.6 for the FE formulation of these equations.

The axisymmetric approach is taken in Solheim and Peltier [1256] (1994), Solheim and Peltier [1257] (1994), and Butler and Jarvis [210] (2004). Check also Thomas and Squyres [1354] (1988), Thomas and Squyres [1353] (1990), Thomas, Squyres, and Carr [1355] (1990), Olson, Schubert, and Anderson [1015] (1993), Schubert, Anderson, and Goldman [1217] (1995).

---

[17]I have not thoroughly checked these equations yet

In cylindrical coordinates

The velocity vector is $\vec{v} = (v_r, v_\theta, v_z)$. Due to the symmetry we have $v_\theta = 0$, $\partial_\theta \to 0$ and the Stokes equations then become [18]

$$-\frac{\partial p}{\partial r} + \eta \left(\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial v_r}{\partial r}) + \frac{\partial^2 v_r}{\partial z^2} - \frac{v_r}{r^2} \right) + \rho g_r \quad = \quad 0 \tag{3.102}$$

$$-\frac{\partial p}{\partial z} + \eta \left(\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial v_z}{\partial r}) + \frac{\partial^2 v_z}{\partial z^2} \right) + \rho g_z \quad = \quad 0 \tag{3.103}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{\partial v_z}{\partial z} \quad = \quad 0 \tag{3.104}$$

The strain rate tensor in cylindrical coordinates is given by

$$\dot{\varepsilon}_{rr} \quad = \quad \frac{\partial v_r}{\partial r} \tag{3.105}$$

$$\dot{\varepsilon}_{\theta\theta} \quad = \quad \frac{v_r}{r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} \tag{3.106}$$

$$\dot{\varepsilon}_{\theta r} = \dot{\varepsilon}_{r\theta} \quad = \quad \frac{1}{2}\left(\frac{\partial v_\theta}{\partial r} - \frac{v_\theta}{r} + \frac{1}{r}\frac{\partial v_r}{\partial \theta} \right) \tag{3.107}$$

$$\dot{\varepsilon}_{zz} \quad = \quad \frac{\partial v_z}{\partial z} \tag{3.108}$$

$$\dot{\varepsilon}_{rz} = \dot{\varepsilon}_{zr} \quad = \quad \frac{1}{2}\left(\frac{\partial v_r}{\partial z} + \frac{\partial v_z}{\partial r} \right) \tag{3.109}$$

$$\dot{\varepsilon}_{\theta z} = \dot{\varepsilon}_{z\theta} \quad = \quad \frac{1}{2}\left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} + \frac{\partial v_\theta}{\partial z} \right) \tag{3.110}$$

In the axisymmetric case, we have $v_\theta = 0$ and $\partial_\theta \to 0$ so that

$$\dot{\varepsilon}_{rr} \quad = \quad \frac{\partial v_r}{\partial r} \tag{3.111}$$

$$\dot{\varepsilon}_{\theta\theta} \quad = \quad \frac{v_r}{r} \tag{3.112}$$

$$\dot{\varepsilon}_{r\theta} = \dot{\varepsilon}_{\theta r} \quad = \quad 0 \tag{3.113}$$

$$\dot{\varepsilon}_{zz} \quad = \quad \frac{\partial v_z}{\partial z} \tag{3.114}$$

$$\dot{\varepsilon}_{rz} = \dot{\varepsilon}_{zr} \quad = \quad \frac{1}{2}\left(\frac{\partial v_r}{\partial z} + \frac{\partial v_z}{\partial r} \right) \tag{3.115}$$

$$\dot{\varepsilon}_{\theta z} = \dot{\varepsilon}_{z\theta} \quad = \quad 0 \tag{3.116}$$

or,

$$\dot{\boldsymbol{\varepsilon}} = \begin{pmatrix} \dot{\varepsilon}_{rr} & 0 & \dot{\varepsilon}_{rz} \\ 0 & \dot{\varepsilon}_{\theta\theta} & 0 \\ \dot{\varepsilon}_{zr} & 0 & \dot{\varepsilon}_{zz} \end{pmatrix}$$

📙 Relevant Literature: Daly & Raefsky (1985) [319], Kiefer & Hager (1992) [741]. This is implemented in STONE 36,63,90,91,92,96,106.

---

[18] https://en.wikipedia.org/wiki/Navier-Stokes_equations

In spherical coordinates

Assuming the flow velocity does not depend on ϕ ($\partial_\phi = 0$) and therefore also that $\mathbf{v}_\phi = 0$

$$0 = -\frac{\partial p}{\partial r} + f_r + \eta\left(\Delta v_r - \frac{2v_r}{r^2} - \frac{2}{r^2}\frac{\partial v_\theta}{\partial\theta} - \frac{2v_\theta\cot\theta}{r^2}\right)$$

$$0 = -\frac{1}{r}\frac{\partial p}{\partial\theta} + \eta\left(\Delta v_\theta + \frac{2}{r^2}\frac{\partial v_r}{\partial\theta} - \frac{v_\theta}{r^2\sin^2\theta}\right)$$

with

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)$$

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}$$

THESE EQUATIONS SHOULD BE CHECKED and RE-CHECKED !!
From [1502]:

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\mathbf{v}_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\mathbf{v}_\theta\sin\theta) + \frac{1}{r\sin\theta}\frac{\partial\mathbf{v}_\phi}{\partial\phi} = 0 \tag{3.117}$$

Pb with 1/r2 ??

$$0 = -\frac{\partial p}{\partial r} + (1-\zeta)Ra\,r\,T + \frac{1}{r^2}\frac{\partial}{\partial r}\left(2\eta r^2\frac{\partial\mathbf{v}_r}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\eta\sin\theta\frac{\partial\mathbf{v}_r}{\partial\theta}\right) + \frac{\partial}{\partial\theta}\left(\eta\frac{\partial}{\partial r}\frac{\mathbf{v}_\theta}{r}\right) \tag{3.118}$$

where $\zeta = R_i/R_o$

The dimensional form of the energy equation in a spherical axisymmetric geometry is given by (assuming the conductivity k to be constant):

$$\rho C_p\left(\frac{\partial T}{\partial t} + \mathbf{v}_r\frac{\partial T}{\partial r} + \frac{\mathbf{v}_\theta}{r}\frac{\partial T}{\partial\theta}\right) = k\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial T}{\partial r}\right) + k\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial T}{\partial\theta}\right)\ldots$$

THESE EQUATIONS SHOULD BE CHECKED and RE-CHECKED !!

3.15 The Boussinesq approximation

As nicely explained in Spiegel & Veronis [1266]: "In the study of problems of thermal convection it is a frequent practice to simplify the basic equations by introducing certain approximations which are attributed to Boussinesq (1903). The Boussinesq approximations can best be summarized by two statements:

1. The fluctuations in density which appear with the advent of motion result principally from thermal (as opposed to pressure) effects.

2. In the equations for the rate of change of momentum and mass, density variations may be neglected except when they are coupled to the gravitational acceleration in the buoyancy force."

Note that their paper examines the Boussinesq approximation for compressible fluids.

[from ASPECT manual] The Boussinesq approximation assumes that the density can be considered constant in all occurrences in the equations with the exception of the buoyancy term on the right hand side of (3.56). The primary result of this assumption is that the continuity equation (3.57) will now read $\vec{\nabla} \cdot \vec{v} = 0$. This implies that the strain rate tensor is deviatoric. Under the Boussinesq approximation, the equations are much simplified:

$$-\vec{\nabla} \cdot [2\eta \dot{\boldsymbol{\varepsilon}}(\vec{v})] + \vec{\nabla} p = \rho \vec{g} \qquad \text{in } \Omega, \qquad (3.119)$$

$$\vec{\nabla} \cdot (\rho \vec{v}) = 0 \qquad \text{in } \Omega, \qquad (3.120)$$

$$\rho_0 C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) - \vec{\nabla} \cdot k \vec{\nabla} T = \rho H \qquad \text{in } \Omega \qquad (3.121)$$

Note that all terms on the rhs of the temperature equations have disappeared, with the exception of the source term.

3.16 The Extended Boussinesq approximation

Yuen *et al.* (2007) [1498] state that the background of the extended Boussinesq equations can be found described in Christensen and Yuen (1985) [269] and more completely in Matyska and Yuen (2007) [891].

📚 Relevant Literature[577, 576]

3.17 Stokes equation for elastic medium

What follows is mostly borrowed from Becker & Kaus lecture notes [71].

The strong form of the PDE that governs force balance in a medium is given by

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{f} = \vec{0}$$

where $\boldsymbol{\sigma}$ is the stress tensor and $\vec{f}$ is a body force.

The stress tensor is related to the strain tensor through the generalised Hooke's law[19]:

$$\sigma_{ij} = \sum_{kl} C_{ijkl} \varepsilon_{kl} \qquad \text{or} \qquad \boldsymbol{\sigma} = \boldsymbol{C} : \boldsymbol{\varepsilon} \tag{3.122}$$

where $\boldsymbol{C}$ is the fourth-order elastic tensor.

Due to the inherent symmetries of $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{C}$, only 21 elastic coefficients of the latter are independent. For isotropic linear media (which have the same physical properties in any direction), $\boldsymbol{C}$ can be reduced to only two independent numbers (for example the bulk modulus K and the shear modulus G that quantify the material's resistance to changes in volume and to shearing deformations, respectively). Thus

$$C_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$

so that Eq. (3.122) becomes:

$$\sigma_{ij} = \lambda \varepsilon_{kk} \delta_{ij} + 2\mu \varepsilon_{ij}$$

or

$$\boldsymbol{\sigma} = \lambda(\vec{\nabla} \cdot \vec{u})\mathbf{1} + 2\mu \boldsymbol{\varepsilon}(\vec{u}) \tag{3.123}$$

where λ is the Lamé parameter and μ is the shear modulus[20]. The term $\vec{\nabla} \cdot \vec{u}$ is the isotropic dilation.

This can be re-written in the 6-dimensional stress/strain space as

$$\underbrace{\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix}}_{\vec{\sigma}} = \underbrace{\begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}}_{C} \cdot \underbrace{\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix}}_{\vec{\varepsilon}}$$

or, in terms of the compliance matrix $\boldsymbol{C}^{-1}$,

$$\vec{\varepsilon} = \boldsymbol{C}^{-1} \cdot \vec{\sigma}$$

with

$$\boldsymbol{C}^{-1} = \frac{1}{\mu(3\lambda+2\mu)} \begin{pmatrix} \lambda+\mu & -\lambda/2 & -\lambda/2 & 0 & 0 & 0 \\ -\lambda/2 & \lambda+\mu & -\lambda/2 & 0 & 0 & 0 \\ -\lambda/2 & -\lambda/2 & \lambda+\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 3\lambda+2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 3\lambda+2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 3\lambda+2\mu \end{pmatrix}$$

---

[19] https://en.wikipedia.org/wiki/Hooke's_law

[20] It is also sometimes written G

If we define the Young's modulus as $E = \mu(3\lambda + 2\mu)/(\lambda + \mu)$ and the Poisson's ratio as $\nu = \lambda(\lambda + \mu)/2$, then

$$\boldsymbol{C}^{-1} = \frac{1}{E} \begin{pmatrix} 1 & -\nu & -\nu & 0 & 0 & 0 \\ -\nu & 1 & -\nu & 0 & 0 & 0 \\ -\nu & -\nu & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2(1+\nu) & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(1+\nu) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(1+\nu) \end{pmatrix}$$

Note that the determinant of $\boldsymbol{C}^{-1}$ is $8(1+\nu)^5(1-2\nu)E^{-6}$, so that when $\nu \to 1/2$ (incompressible material), the compliance matrix is singular and the stress cannot be given as a function of strain [862].

The strain tensor is related to the displacement as follows:

$$\boldsymbol{\varepsilon}(\vec{u}) = \frac{1}{2}(\vec{\nabla}\vec{u} + (\vec{\nabla}\vec{u})^T)$$

The incompressibility (or bulk modulus) K is defined as $p = -K\vec{\nabla} \cdot \vec{u}$ where p is the pressure with

$$\begin{aligned} p &= -\frac{1}{3}\text{tr}(\boldsymbol{\sigma}) \\ &= -\frac{1}{3}[\lambda(\vec{\nabla} \cdot \vec{u})\text{tr}[\mathbf{1}] + 2\mu\text{tr}[\boldsymbol{\varepsilon}(\vec{u})]] \\ &= -\frac{1}{3}[\lambda(\vec{\nabla} \cdot \vec{u})3 + 2\mu(\vec{\nabla} \cdot \vec{u})] \\ &= -\left[\lambda + \frac{2}{3}\mu\right](\vec{\nabla} \cdot \vec{u}) \end{aligned} \tag{3.124}$$

so that

$$p = -K\vec{\nabla} \cdot \vec{u} \qquad \text{with} \qquad K = \lambda + \frac{2}{3}\mu$$

Remark. *Eq. (3.122) and (3.123) are analogous to the ones that one has to solve in the context of viscous flow using the penalty method. In this case λ is the penalty coefficient, $\vec{u}$ is the velocity, and μ is then the dynamic viscosity.*

The Lamé parameter and the shear modulus are also linked to ν the poisson ratio, and E, Young's modulus:

$$\lambda = \mu\frac{2\nu}{1 - 2\nu} = \frac{\nu E}{(1+\nu)(1-2\nu)} \qquad \text{with} \qquad E = 2\mu(1+\nu)$$

The shear modulus, expressed often in GPa, describes the material's response to shear stress. The poisson ratio describes the response in the direction orthogonal to uniaxial stress. The Young modulus, expressed in GPa, describes the material's strain response to uniaxial stress in the direction of this stress.

📥 Relevant Literature: solvers for 3D Stokes and elasticity problems with heterogeneous coefficients [1182]

3.18 Boundary conditions

In mathematics, the Dirichlet (or first-type) boundary condition is a type of boundary condition, named after Peter Gustav Lejeune Dirichlet. When imposed on an ODE or PDE, it specifies the values that a solution needs to take along the boundary of the domain. Note that a Dirichlet boundary condition may also be referred to as a fixed boundary condition.

The Neumann (or second-type) boundary condition is a type of boundary condition, named after Carl Neumann. When imposed on an ordinary or a partial differential equation, the condition specifies the values in which the derivative of a solution is applied within the boundary of the domain.

It is possible to describe the problem using other boundary conditions: a Dirichlet boundary condition specifies the values of the solution itself (as opposed to its derivative) on the boundary, whereas the Cauchy boundary condition, mixed boundary condition and Robin boundary condition are all different types of combinations of the Neumann and Dirichlet boundary conditions.

3.18.1 The Stokes equations

You may find the following terms in the computational geodynamics literature:

- free surface: this means that no force is acting on the surface, i.e. $\boldsymbol{\sigma} \cdot \vec{n} = \vec{0}$. It is usually used on the top boundary of the domain and allows for topography evolution.

- free slip: $\vec{v} \cdot \vec{n} = 0$ and $(\boldsymbol{\sigma} \cdot \vec{n}) \times \vec{n} = \vec{0}$. This condition ensures a frictionless flow parallel to the boundary where it is prescribed.

- no slip: this means that the velocity (or displacement) is exactly zero on the boundary, i.e. $\vec{v} = \vec{0}$.

- prescribed velocity: $\vec{v} = \vec{v}_{bc}$

- stress b.c.:

- open .b.c.: see STONE 29.

3.18.2 The heat transport equation

There are two types of boundary conditions for this equation: temperature boundary conditions (Dirichlet boundary conditions) and heat flux boundary conditions (Neumann boundary conditions).

3.19 Meaningful physical quantities

- Velocity $\vec{v}$(m/s): This is a vector quantity and both magnitude and direction are needed to define it. It is the rate of change of position with respect to a frame of reference.

- Root mean square velocity ν_{rms}(m/s):

$$\nu_{rms} = \left(\frac{\int_\Omega |\vec{v}|^2 \, dV}{\int_\Omega dV} \right)^{1/2} = \left(\frac{1}{V_\Omega} \int_\Omega |\vec{v}|^2 \, dV \right)^{1/2} \tag{3.125}$$

Remark. *V_Ω is usually computed numerically at the same time that ν_{vrms} is computed.*

In Cartesian coordinates, for a cuboid domain of size $Lx \times L_y \times Lz$, the ν_{rms} is simply given by:

$$\nu_{rms} = \left(\frac{1}{L_x L_y L_z} \int_0^{L_x} \int_0^{L_y} \int_0^{L_z} (u^2 + v^2 + w^2) dx dy dz \right)^{1/2} \tag{3.126}$$

In the case of an annulus domain, although calculations are carried out in Cartesian coordinates, it makes sense to look at the radial velocity component ν_r and the tangential velocity component ν_θ, and their respective root mean square averages:

$$\nu_r|_{rms} = \left(\frac{1}{V_\Omega} \int_\Omega v_r^2 \, d\Omega \right)^{1/2} \tag{3.127}$$

$$\nu_\theta|_{rms} = \left(\frac{1}{V_\Omega} \int_\Omega v_\theta^2 \, d\Omega \right)^{1/2} \tag{3.128}$$

- Pressure p (Pa):

- Stress tensor $\boldsymbol{\sigma}$ (Pa):

- Strain tensor $\boldsymbol{\varepsilon}$ (dimensionless):

- Strain rate tensor $\dot{\boldsymbol{\varepsilon}}$ (s^{-1}):

- Argand Number: Non-dimensional number (Ar) representing the ratio of the stress arising from crustal thickness contrasts (vertical stress) to the stress required to deform the material at ambient strain rates (horizontal stress) It is commonly used in mountain building dynamics as a measure of the tendency of an orogen to collapse under its own gravitational potential energy. See England & McKenzie [399], Houseman & England [629].

- (Thermal) Rayleigh number Ra (or Ra$_T$) (X): It is a dimensionless number that expresses the ratio of the driving forces to the opposing forces. The buoyancy force comes from the volumetric thermal expansion while the viscous forces and the heat diffusivity oppose convection (the latter one smoothes out thermal gradients).

The Rayleigh number for convection driven by a constant temperature hot base and a cold surface in a domain of thickness D is:

$$\mathsf{Ra} = \frac{\rho_0 g \alpha D^3}{\eta \kappa} \cdot \Delta T = \frac{\rho_0^2 C_p g \alpha D^3 \Delta T}{\eta k}$$

The Rayleigh number for convection driven by a hot base (constant basal heat flow q_b) and a colder surface is:

$$\mathsf{Ra} = \frac{\rho_0 g \alpha D^3}{\eta \kappa} \cdot \frac{q_b D}{k}$$

The Rayleigh number for convection driven by internal heating H (production per cubic meter) is:

$$\mathsf{Ra} = \frac{\rho_0 g \alpha D^3}{\eta \kappa} \cdot \frac{H D^2}{k}$$

The Rayleigh number for convection driven by both basal heat flow and internal heating is:

$$\mathsf{Ra} = \frac{\rho_0 g \alpha D^3}{\eta \kappa} \cdot \frac{q_b D + H D^2}{k}$$

For convection to occur, the Rayleigh number must be larger than the so-called critical Rayleigh number, which ranges from 600 to 3000 (it depends on the boundary conditions and the geometry).

- Compositional Rayleigh Number Ra_C:

$$\mathsf{Ra}_C = \frac{\Delta \rho_C g D^3}{\kappa \eta_0}$$

where $\Delta \rho_C$ is the difference in density between the distinct material compositions (when compared at identical temperatures). See for instance Trim *et al.* (2020) [1371].

- Prandtl number Pr (X): It is named after the German physicist Ludwig Prandtl[21] and is defined as the ratio of momentum diffusivity to thermal diffusivity. It is given as:

$$\mathsf{Pr} = \frac{\text{momentum diffusivity}}{\text{thermal diffusivity}} = \frac{\eta/\rho}{k/(\rho C_p)} = \frac{\eta C_p}{k}$$

For Earth materials, we have $Pr \sim (10^{21} 1000)/3 >> 1$, which means that momentum diffusivity dominates.

- Nusselt number Nu (X): the Nusselt number (Nu) is the ratio of convective to conductive heat transfer across (normal to) the boundary. The conductive component is measured under the same conditions as the heat convection but with a (hypothetically) stagnant (or motionless) fluid.

In practice the Nusselt number Nu of a layer (typically the mantle of a planet) is defined as follows:

$$\mathsf{Nu} = \frac{q}{q_c} \tag{3.129}$$

where q is the heat transferred by convection while $q_c = k \Delta T / D$ is the amount of heat that would be conducted through a layer of thickness D with a temperature difference ΔT across it with k being the thermal conductivity.

For 2D Cartesian systems of size (L_x, L_y) the Nu is computed [102]

$$\mathsf{Nu} = \frac{\frac{1}{L_x}\int_0^{L_x} k \frac{\partial T}{\partial y}(x, y = L_y)dx}{-\frac{1}{L_x}\int_0^{L_x} k T(x, y = 0)/L_y dx} = -L_y \frac{\int_0^{L_x} \frac{\partial T}{\partial y}(x, y = L_y)dx}{\int_0^{L_x} T(x, y = 0)dx}$$

i.e. it is the mean surface temperature gradient over the mean bottom temperature.

---

[21] https://en.wikipedia.org/wiki/Ludwig_Prandtl

Note that in the case when no convection takes place then the measured heat flux at the top is the one obtained from a purely conductive profile which yields Nu=1.

Note that a relationship $\text{Ra} \propto \text{Nu}^\alpha$ exists between the Rayleigh number Ra and the Nusselt number Nu in convective systems, see [1468] and references therein.

Turning now to cylindrical geometries with inner radius R_1 and outer radius R_2, we define $f = R_1/R_2$. A small value of f corresponds to a high degree of curvature. We assume now that $R_2 - R_1 = 1$, so that $R_2 = 1/(1-f)$ and $R_1 = f/(1-f)$. Following [669], the Nusselt number at the inner and outer boundaries are:

$$\boxed{\text{Nu}_{inner} = \frac{f \ln f}{1-f} \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\partial T}{\partial r} \right)_{r=R_1} d\theta} \tag{3.130}$$

$$\boxed{\text{Nu}_{outer} = \frac{\ln f}{1-f} \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\partial T}{\partial r} \right)_{r=R_2} d\theta} \tag{3.131}$$

Note that a conductive geotherm in such an annulus between temperatures T_1 and T_2 is given by

$$T_c(r) = \frac{\ln(r/R_2)}{\ln(R_1/R_2)} = \frac{\ln(r(1-f))}{\ln f}$$

so that

$$\frac{\partial T_c}{\partial r} = \frac{1}{r} \frac{1}{\ln f}$$

We then find:

$$\text{Nu}_{inner} = \frac{f \ln f}{1-f} \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\partial T_c}{\partial r} \right)_{r=R_1} d\theta = \frac{f \ln f}{1-f} \frac{1}{R_1} \frac{1}{\ln f} = 1 \tag{3.132}$$

$$\text{Nu}_{outer} = \frac{\ln f}{1-f} \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\partial T_c}{\partial r} \right)_{r=R_2} d\theta = \frac{\ln f}{1-f} \frac{1}{R_2} \frac{1}{\ln f} = 1 \tag{3.133}$$

As expected, the recovered Nusselt number at both boundaries is exactly 1 when the temperature field is given by a steady state conductive geotherm.

📚 Relevant Literature[846]

- Temperature (K):

- (Dynamic) Viscosity (Pa s): For air it is roughly 10^{-5}Pa s, about 10^{-3} Pa s for water, about 10^{10} Pa s for ice and about 10^{17} Pa s for salt.

- Entropy S (J K^{-1})

- (mass) Density ρ (kg m^{-3}):

- Heat capacity C_p (J K^{-1}): It is the measure of the heat/energy required to increase the temperature of a unit quantity of a substance by unit degree. Note that the *specific* heat capacity c_P of a substance is the heat capacity of a sample of the substance divided by the mass of the sample, with units $\text{J K}^{-1} \text{kg}^{-1}$.

"Different substances respond to heat in different ways. If a metal chair sits in the bright sun on a hot day, it may become quite hot to the touch. An equal mass of water under the same

sun exposure will not become nearly as hot. This means that water has a high heat capacity (the amount of heat required to raise the temperature of an object by 1 °C). Water is very resistant to changes in temperature, while metals generally are not." [22]

- **Heat conductivity**, or thermal conductivity k ($\text{W m}^{-1}\text{K}^{-1}$). It is the property of a material that indicates its ability to conduct heat. It appears primarily in Fourier's Law for heat conduction. Note that it is a function of temperature, especially in mantle convection settings, see Bonneville & Capolsini (1999) [122] and refs therein, Miyauchi & Kameyama (2013) [936], Hofmeister & Yuen (2007) [616]. Note also that it can be a tensorial quantity in anisotropic context. The heat conductivity of many rocks was determined in [23].

- **Heat diffusivity**: $\kappa = k/(\rho C_p)$ (m^2s^{-1}). Substances with high thermal diffusivity rapidly adjust their temperature to that of their surroundings, because they conduct heat quickly in comparison to their volumetric heat capacity or 'thermal bulk'.

- **thermal expansion** α (K^{-1}): it is the tendency of a matter to change in volume in response to a change in temperature. Note that it is a function of temperature, especially in mantle convection settings [936].

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P$$

- **Urey Ratio**: mantle heat production divided by heat loss. It is a key constraint for thermal history models. Recent Urey ratio estimates are in the range of 0.21-0.49. [814]

- **Shear modulus**: modulus of rigidity, usually expressed in GPa. It describes the material response to shear stress.

- **Poisson ratio**: response in the direction orthogonal to uniaxial stress.

- **Young's modulus**: describes the material strain response to uniaxial stress in the direction of this stress, usually expressed in GPa.

- **Average viscosity**: following Christensen (1983) [260], one can compute the averaged viscosity in a domain as follows:

$$\langle \eta \rangle = \frac{\int_V \eta \dot{\varepsilon}_e^2 dV}{\int_V \dot{\varepsilon}_e^2 dV} \tag{3.134}$$

check aspect manual The 2D cylindrical shell benchmarks by Davies *et al.* 5.4.12

---

[22] https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introductory_Chemistry_(CK-12)
/17%3A_Thermochemistry/17.04%3A_Heat_Capacity_and_Specific_Heat

3.20 Moment of inertia

Consider a rigid body rotating with fixed angular velocity ω about an axis which passes through the origin. Let $\boldsymbol{r}_i$ be the position vector of the ith mass element, whose mass is m_i. We expect this position vector to precess about the axis of rotation (which is parallel to ω) with angular velocity ω.

$$\frac{d\boldsymbol{r}_i}{dt} = \boldsymbol{\omega} \times \boldsymbol{r}_i.$$

Thus, the above equation specifies the velocity, $\boldsymbol{v}_i = d\boldsymbol{r}_i/dt$, of each mass element as the body rotates with fixed angular velocity ω about an axis passing through the origin.

The total angular momentum of the body (about the origin) is written

$$\boldsymbol{L} = \sum_{i=1,N} m_i\,\boldsymbol{r}_i \times \frac{d\boldsymbol{r}_i}{dt} = \sum_{i=1,N} m_i\,\boldsymbol{r}_i \times (\boldsymbol{\omega} \times \boldsymbol{r}_i) = \sum_{i=1,N} m_i\,[r_i^2\boldsymbol{\omega} - (\boldsymbol{r}_i \cdot \boldsymbol{\omega})\boldsymbol{r}_i]$$

The above formula can be written as a matrix equation of the form

$$\begin{pmatrix} L_x \\ L_y \\ L_z \end{pmatrix} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix}$$

where

$$I_{xx} = +\sum_{i=1,N} (y_i^2 + z_i^2)\,m_i = \int (y^2 + z^2)\,dm = \int_V (y^2 + z^2)\,\rho(x,y,z)dV$$

$$I_{yy} = +\sum_{i=1,N} (x_i^2 + z_i^2)\,m_i = \int (x^2 + z^2)\,dm = \int_V (x^2 + z^2)\,\rho(x,y,z)dV$$

$$I_{zz} = +\sum_{i=1,N} (x_i^2 + y_i^2)\,m_i = \int (x^2 + y^2)\,dm = \int_V (x^2 + y^2)\,\rho(x,y,z)dV$$

$$I_{xy} = I_{yx} = -\sum_{i=1,N} x_i\,y_i\,m_i = -\int x\,y\,dm = -\int x\,y\,\rho(x,y,z)dV$$

$$I_{yz} = I_{zy} = -\sum_{i=1,N} y_i\,z_i\,m_i = -\int y\,z\,dm = -\int y\,z\,\rho(x,y,z)dV$$

$$I_{xz} = I_{zx} = -\sum_{i=1,N} x_i\,z_i\,m_i = -\int x\,z\,dm = -\int x\,z\,\rho(x,y,z)dV$$

Here, I_{xx} is called the moment of inertia about the x-axis, I_{yy} the moment of inertia about the y-axis, I_{xy} the xy product of inertia, I_{yz} the yz product of inertia, etc. The matrix of the I_{ij} values is known as the moment of inertia tensor.

In general, the angular momentum vector, $\mathbf{L}$ points in a different direction to the angular velocity vector, ω. In other words, $\mathbf{L}$ is generally not parallel to ω.

Finally, although the above results were obtained assuming a fixed angular velocity, they remain valid at each instant in time if the angular velocity varies.

In the simplified case of a spherically symmetric planet, it is easy to see that $I_{xx} = I_{yy} = I_{zz}$ so that $I = \frac{1}{3}(I_{xx} + I_{yy} + I_{zz})$, and $\rho = \rho(r)$ with $dV = 4\pi r^2 dr$, leading to

$$I = \frac{8\pi}{3}\int_0^R \rho(r)r^4 dr$$

Assuming further that the planet has a constant density ρ_0, we obtain

$$I = \frac{8\pi}{3}\rho_0 \int_0^R r^4 dr = \frac{8\pi}{3}\rho_0 \frac{R^5}{5} = \frac{2}{5}MR^2$$

where M is the mass of the planet and R is its radius.

Assuming now that the planet is composed of a core of radius R_c and density ρ_c surrounded by a mantle of density ρ_m, we have

$$I = \frac{8\pi}{3}\int_0^R \rho(r)r^4 dr = \frac{8\pi}{3}\left(\int_0^{R_c}\rho_c r^4 dr + \int_{R_c}^R \rho_m r^4 dr\right) = \frac{8\pi}{15}\left(\rho_c R_c^5 + \rho_m(R^5 - R_c^5)\right)$$

The moment of inertia of the core is given in Table 2 of "Core Dynamics", Treatise on Geophysics, edited by Peter Olson: $I_{core} = 9.2 \times 10^{36} kg.m^2$. The total moment of inertia for the Earth is then given by $I = I_{core} + I_{mantle}$.

Chapter 4

Tensors, coordinate systems and invariants

4.1 The strain rate tensor in all coordinate systems

The strain rate tensor $\dot{\boldsymbol{\varepsilon}}(\vec{v})$ is given by

$$\dot{\boldsymbol{\varepsilon}}(\vec{v}) = \frac{1}{2} \left(\vec{\nabla}\vec{v} + (\vec{\nabla}\vec{v})^T \right) \tag{4.1}$$

4.1.1 Cartesian coordinates

$$\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} \tag{4.2}$$

$$\dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} \tag{4.3}$$

$$\dot{\varepsilon}_{zz} = \frac{\partial w}{\partial z} \tag{4.4}$$

$$\dot{\varepsilon}_{yx} = \dot{\varepsilon}_{xy} = \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \tag{4.5}$$

$$\dot{\varepsilon}_{zx} = \dot{\varepsilon}_{xz} = \frac{1}{2} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right) \tag{4.6}$$

$$\dot{\varepsilon}_{zy} = \dot{\varepsilon}_{yz} = \frac{1}{2} \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right) \tag{4.7}$$

In the ASPECT manual there is an interesting discussion about the strain rate tensor in the case of 2D models:

> The notion we adopt here is to think of two-dimensional models in the following way: We assume that the domain we want to solve on is a two-dimensional cross section (parameterized by x and y coordinates) that extends infinitely far in both negative and positive z direction. Further, we assume that the velocity is zero in z direction and that all variables have no variation in z direction. As a consequence, we ought to really think of these two-dimensional models as three-dimensional ones in which the z component of the velocity is zero and so are all z derivatives.

This of course makes sense but it means that when the deviatoric strain rate tensor needs to be computed, then it is given by

$$\dot{\boldsymbol{\varepsilon}}^d = \dot{\boldsymbol{\varepsilon}} - \frac{1}{3}(\vec{\nabla}\cdot\vec{v})\mathbf{1} = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & 0 \\ \dot{\varepsilon}_{xy} & \dot{\varepsilon}_{yy} & 0 \\ 0 & 0 & 0 \end{pmatrix} - \frac{1}{3}(\dot{\varepsilon}_{xx}+\dot{\varepsilon}_{yy})\mathbf{1} = \frac{1}{3}\begin{pmatrix} 2\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} & 3\dot{\varepsilon}_{xy} & 0 \\ 3\dot{\varepsilon}_{xy} & -\dot{\varepsilon}_{xx} + 2\dot{\varepsilon}_{yy} & 0 \\ 0 & 0 & -\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} \end{pmatrix}$$

As a consequence the shear heating term Φ is given by

$$
\begin{aligned}
\Phi = 2\eta\dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d &= 2\eta\frac{1}{9}\left[(2\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + (-\dot{\varepsilon}_{xx} + 2\dot{\varepsilon}_{yy})^2 + 2\cdot 9\dot{\varepsilon}_{xy}^2 + (-\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2\right] \\
&= 2\eta\frac{1}{9}\left[4\dot{\varepsilon}_{xx}^2 - 4\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + \dot{\varepsilon}_{yy}^2 + \dot{\varepsilon}_{xx}^2 - 4\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + 4\dot{\varepsilon}_{yy}^2 + 18\dot{\varepsilon}_{xy}^2 + \dot{\varepsilon}_{xx}^2 + 2\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + \dot{\varepsilon}_{yy}^2\right] \\
&= 2\eta\frac{1}{9}\left[6\dot{\varepsilon}_{xx}^2 + 6\dot{\varepsilon}_{yy}^2 - 6\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + 18\dot{\varepsilon}_{xy}^2\right] \\
&= 2\eta\left[\frac{2}{3}\dot{\varepsilon}_{xx}^2 + \frac{2}{3}\dot{\varepsilon}_{yy}^2 - \frac{2}{3}\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{xy}^2\right] \quad (4.8)
\end{aligned}
$$

4.1.2 Polar coordinates

$$
\begin{aligned}
\dot{\varepsilon}_{rr} &= \frac{\partial v_r}{\partial r} \\
\dot{\varepsilon}_{\theta\theta} &= \frac{v_r}{r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} \\
\dot{\varepsilon}_{\theta r} = \dot{\varepsilon}_{r\theta} &= \frac{1}{2}\left(\frac{\partial v_\theta}{\partial r} - \frac{v_\theta}{r} + \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)
\end{aligned}
$$

4.1.3 Cylindrical coordinates

$$\dot{\varepsilon}_{rr} = \frac{\partial v_r}{\partial r} \tag{4.9}$$

$$\dot{\varepsilon}_{\theta\theta} = \frac{v_r}{r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} \tag{4.10}$$

$$\dot{\varepsilon}_{\theta r} = \dot{\varepsilon}_{r\theta} = \frac{1}{2}\left(\frac{\partial v_\theta}{\partial r} - \frac{v_\theta}{r} + \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right) \tag{4.11}$$

$$\dot{\varepsilon}_{zz} = \frac{\partial v_z}{\partial z} \tag{4.12}$$

$$\dot{\varepsilon}_{rz} = \dot{\varepsilon}_{zr} = \frac{1}{2}\left(\frac{\partial v_r}{\partial z} + \frac{\partial v_z}{\partial r}\right) \tag{4.13}$$

$$\dot{\varepsilon}_{\theta z} = \dot{\varepsilon}_{z\theta} = \frac{1}{2}\left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} + \frac{\partial v_\theta}{\partial z}\right) \tag{4.14}$$

The velocity divergence is given by

$$\vec{\nabla}\cdot\vec{v} = \dot{\varepsilon}_{rr} + \dot{\varepsilon}_{\theta\theta} + \dot{\varepsilon}_{zz} = \frac{\partial v_r}{\partial r} + \frac{1}{r}\left(\frac{\partial v_\theta}{\partial \theta} + v_r\right) + \frac{\partial v_z}{\partial z} \tag{4.15}$$

4.1.4 Spherical coordinates

$$\dot{\varepsilon}_{rr} = \frac{\partial \nu_r}{\partial r} \tag{4.16}$$

$$\dot{\varepsilon}_{\theta\theta} = \frac{\nu_r}{r} + \frac{1}{r}\frac{\partial \nu_\theta}{\partial \theta} \tag{4.17}$$

$$\dot{\varepsilon}_{\phi\phi} = \frac{1}{r\sin\theta}\frac{\partial \nu_\phi}{\partial \phi} + \frac{\nu_r}{r} + \frac{\nu_\theta \cot\theta}{r} \tag{4.18}$$

$$\dot{\varepsilon}_{\theta r} = \dot{\varepsilon}_{r\theta} = \frac{1}{2}\left(r\frac{\partial}{\partial r}(\frac{\nu_\theta}{r}) + \frac{1}{r}\frac{\partial \nu_r}{\partial \theta}\right) \tag{4.19}$$

$$\dot{\varepsilon}_{\phi r} = \dot{\varepsilon}_{r\phi} = \frac{1}{2}\left(\frac{1}{r\sin\theta}\frac{\partial \nu_r}{\partial \phi} + r\frac{\partial}{\partial r}(\frac{\nu_\phi}{r})\right) \tag{4.20}$$

$$\dot{\varepsilon}_{\phi\theta} = \dot{\varepsilon}_{\theta\phi} = \frac{1}{2}\left(\frac{\sin\theta}{r}\frac{\partial}{\partial \theta}(\frac{\nu_\phi}{\sin\theta}) + \frac{1}{r\sin\theta}\frac{\partial \nu_\theta}{\partial \phi}\right) \tag{4.21}$$

4.1.5 Relationship between Cartesian and polar coordinates expressions

We can go from Cartesian to polar coordinates via the 2×2 transformation matrix:

$$\mathcal{P} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{4.22}$$

The rows correspond to the components of $\vec{e}_r$ and $\vec{e}_\theta$ in the Cartesian basis. A vector $\vec{v}$ transforms from one orthonormal basis to another by multiplying it by the matrix $\mathcal{P}$. As we have seen before, this yields

$$\nu_r = u\cos\theta + v\sin\theta \tag{4.23}$$
$$\nu_\theta = -u\sin\theta + v\cos\theta \tag{4.24}$$

A second-order tensor $\boldsymbol{a}$ is Cartesian coordinates transforms into $\boldsymbol{a}^\star$ in polar coordinates by

$$\boldsymbol{a}^\star = \mathcal{P} \cdot \boldsymbol{a} \cdot \mathcal{P}^T$$

and obviously

$$\boldsymbol{a} = \mathcal{P}^T \cdot \boldsymbol{a}^\star \cdot \mathcal{P}$$

We obtain for the strain rate tensor (or the stress tensor):

$$\dot{\varepsilon}_{rr} = \dot{\varepsilon}_{xx}\cos^2\theta + \dot{\varepsilon}_{yy}\sin^2\theta + 2\dot{\varepsilon}_{xy}\sin\theta\cos\theta$$
$$\dot{\varepsilon}_{\theta\theta} = \dot{\varepsilon}_{xx}\sin^2\theta + \dot{\varepsilon}_{yy}\cos^2\theta - 2\dot{\varepsilon}_{xy}\sin\theta\cos\theta$$
$$\dot{\varepsilon}_{r\theta} = \dot{\varepsilon}_{xy}(\cos^2\theta - \sin^2\theta) + (\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{xx})\sin\theta\cos\theta$$

Using the trigonometric identities $\sin 2\theta = 2\sin\theta\cos\theta$ and $\cos^2\theta - \sin^2\theta = \cos 2\theta$, then we obtain

$$\dot{\varepsilon}_{rr} = \dot{\varepsilon}_{xx}\cos^2\theta + \dot{\varepsilon}_{yy}\sin^2\theta + \dot{\varepsilon}_{xy}\sin 2\theta$$
$$\dot{\varepsilon}_{\theta\theta} = \dot{\varepsilon}_{xx}\sin^2\theta + \dot{\varepsilon}_{yy}\cos^2\theta - \dot{\varepsilon}_{xy}\sin 2\theta$$
$$\dot{\varepsilon}_{r\theta} = \dot{\varepsilon}_{xy}\cos 2\theta + \frac{1}{2}(\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{xx})\sin 2\theta$$

and likewise:

$$\dot{\varepsilon}_{xx} = \dot{\varepsilon}_{rr}\cos^2\theta + \dot{\varepsilon}_{\theta\theta}\sin^2\theta - 2\dot{\varepsilon}_{r\theta}\sin\theta\cos\theta \tag{4.25}$$
$$\dot{\varepsilon}_{yy} = \dot{\varepsilon}_{rr}\sin^2\theta + \dot{\varepsilon}_{\theta\theta}\cos^2\theta + 2\dot{\varepsilon}_{r\theta}\sin\theta\cos\theta \tag{4.26}$$
$$\dot{\varepsilon}_{xy} = \dot{\varepsilon}_{r\theta}(\cos^2\theta - \sin^2\theta) + (\dot{\varepsilon}_{rr} - \dot{\varepsilon}_{\theta\theta})\sin\theta\cos\theta \tag{4.27}$$

4.2 Principal stress and principal invariants

As seen before (see Section 3.4.2) the stress tensor is a symmetric 3×3 real matrix, and linear algebra tells us that it therefore has three mutually orthogonal unit-length eigenvectors $\vec{n}_1$, $\vec{n}_2$, $\vec{n}_3$ and three real eigenvalues $\lambda_1, \lambda_2, \lambda_3$ such that $\boldsymbol{\sigma} \cdot \vec{n}_i = \lambda_i \vec{n}_i$.

As a consequence, in a coordinate system with axes $\vec{n}_1, \vec{n}_2, \vec{n}_3$, the stress tensor is a diagonal matrix, and has only the three normal components $\lambda_1, \lambda_2, \lambda_3$ i.e. the principal stresses. If the three eigenvalues are equal, the stress is an isotropic compression or tension, always perpendicular to any surface, there is no shear stress, and the tensor is a diagonal matrix in any coordinate frame.

4.2.1 In two dimensions

We are looking for the stress tensor eigenvector vector $\vec{n} = (n_x, n_y)$ associated to the eigenvalue λ such that

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} = \lambda \begin{pmatrix} n_x \\ n_y \end{pmatrix}$$

or,

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} = \vec{0}$$

i.e.,

$$\begin{pmatrix} \sigma_{xx} - \lambda & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} - \lambda \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} = \vec{0}$$

which yields

$$(\sigma_{xx} - \lambda)(\sigma_{yy} - \lambda) - \sigma_{xy}^2 = 0$$

or,

$$\lambda^2 - (\sigma_{xx} + \sigma_{yy})\lambda + (\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2) = 0$$

The discriminant Δ is

$$\begin{aligned} \Delta &= (\sigma_{xx} + \sigma_{yy})^2 - 4(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2) \\ &= (\sigma_{xx} - \sigma_{yy})^2 + 4\sigma_{xy}^2 \end{aligned}$$

The roots are given by:

$$\begin{aligned} \lambda_{\pm} &= \frac{(\sigma_{xx} + \sigma_{yy}) \pm \sqrt{(\sigma_{xx} - \sigma_{yy})^2 + 4\sigma_{xy}^2}}{2} \\ &= \frac{\sigma_{xx} + \sigma_{yy}}{2} \pm \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2} \end{aligned}$$

The two principal stresses are then:

$$\begin{aligned} \sigma_1 &= \frac{\sigma_{xx} + \sigma_{yy}}{2} + \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2} \\ \sigma_2 &= \frac{\sigma_{xx} + \sigma_{yy}}{2} - \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2} \end{aligned} \tag{4.28}$$

with the convention $\sigma_1 > \sigma_2$. The maximum shear stress is defined as one-half the difference between the two principal stresses

$$\tau_{max} = \frac{\sigma_1 - \sigma_2}{2} = \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2} \tag{4.29}$$

The eigenvector $\vec{n}_1$ corresponding to σ_1 is obtained by solving

$$\boldsymbol{\sigma} \cdot \vec{n}_1 = \sigma_1 \vec{n}_1$$

and same for the other eigenvalue/vector:

$$\boldsymbol{\sigma} \cdot \vec{n}_2 = \sigma_2 \vec{n}_2$$

Each is a system of two equations with two unknowns. These are not difficult to solve, but can prove cumbersome. Note that linear algebra tells us that $\vec{n}_1 \cdot \vec{n}_2 = 0$, i.e. the eigenvectors form a basis of $\mathbb{R}^2$.

This is the reason why often people go another route. One can ask the question: what is the value of the angle θ_p which, if used to perform a rotation of the axis system, yields a stress tensor that is diagonal, with the principal stresses on the diagonal?

Stresses in given coordinate system

Principal stresses
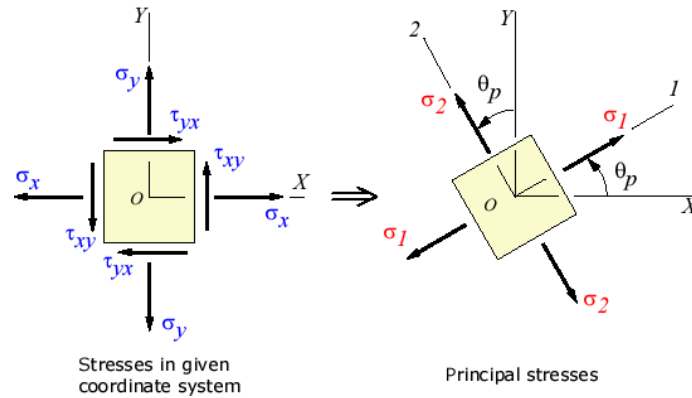
Taken from `https://www.efunda.com/formulae/solid_mechanics/mat_mechanics/plane_stress_principal.cfm`

The rotation matrix is
$$\boldsymbol{R} = \begin{pmatrix} \cos\theta_p & -\sin\theta_p \\ \sin\theta_p & \cos\theta_p \end{pmatrix}$$

and the image of $\boldsymbol{\sigma}$ by means of the axis rotation is $\boldsymbol{\sigma}' = \boldsymbol{R} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{R}^{-1}$, i.e.

$$\begin{aligned}
\boldsymbol{\sigma}' &= \begin{pmatrix} \cos\theta_p & -\sin\theta_p \\ \sin\theta_p & \cos\theta_p \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \cdot \begin{pmatrix} \cos\theta_p & \sin\theta_p \\ -\sin\theta_p & \cos\theta_p \end{pmatrix} \\
&= \begin{pmatrix} \cos\theta_p & -\sin\theta_p \\ \sin\theta_p & \cos\theta_p \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx}\cos\theta_p - \sigma_{xy}\sin\theta_p & \sigma_{xx}\sin\theta_p + \sigma_{xy}\cos\theta_p \\ \sigma_{xy}\cos\theta_p - \sigma_{yy}\sin\theta_p & \sigma_{xy}\sin\theta_p + \sigma_{yy}\cos\theta_p \end{pmatrix} \\
&= \begin{pmatrix} \dots & \cos\theta_p(\sigma_{xx}\sin\theta_p + \sigma_{xy}\cos\theta_p) - \sin\theta_p(\sigma_{xy}\sin\theta_p + \sigma_{yy}\cos\theta_p) \\ \dots & \dots \end{pmatrix}
\end{aligned}$$

In the matrix above I have only computed the off diagonal term since we are actually looking for θ_p such that $\sigma'_{xy} = 0$, or

$$\begin{aligned}
\cos\theta_p(\sigma_{xx}\sin\theta_p + \sigma_{xy}\cos\theta_p) - \sin\theta_p(\sigma_{xy}\sin\theta_p + \sigma_{yy}\cos\theta_p) &= 0 \\
\sin\theta_p\cos\theta_p(\sigma_{xx} - \sigma_{yy}) + (\cos^2\theta_p - \sin^2\theta_p)\sigma_{xy} &= 0
\end{aligned}$$

and then

$$\frac{\sin\theta_p \cos\theta_p}{\cos^2\theta_p - \sin^2\theta_p} = \frac{\sigma_{xy}}{\sigma_{xx} - \sigma_{yy}}$$

The left hand term is actually a trigonometric identity[1]:

$$\frac{\sin\theta_p \cos\theta_p}{\cos^2\theta_p - \sin^2\theta_p} = \frac{\frac{1}{2}\sin 2\theta_p}{\cos 2\theta_p} = \frac{1}{2}\tan 2\theta_p$$

and finally:

$$\tan 2\theta_p = \frac{2\sigma_{xy}}{\sigma_{xx} - \sigma_{yy}} \qquad \text{or} \qquad \boxed{\theta_p = \frac{1}{2}\tan^{-1}\frac{2\sigma_{xy}}{\sigma_{xx} - \sigma_{yy}}}$$

Once θ_p has been found the other direction is given by $\theta_p + \pi/2$.

Example: Let us assume a diagonal stress tensor of the form

$$\boldsymbol{\sigma} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

then $\tan 2\theta_p = 0$, and then $\theta_p = 0$. The principal directions are the horizontal and vertical directions, i.e. the Cartesian axis system, which is consistent.

> add a remark that 2D does not exist and that plane strain incompressible actually is what is going on above

4.2.2 In three dimensions

We are looking for the stress tensor eigenvector vector $\vec{n} = (n_x, n_y, n_z)$ associated to the eigenvalue λ such that

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \lambda \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix}$$

or,

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \vec{0}$$

$$\begin{pmatrix} \sigma_{xx} - \lambda & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} - \lambda & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} - \lambda \end{pmatrix} \cdot \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \vec{0}$$

Non-trivial solutions of this equation require

$$\begin{vmatrix} \sigma_{xx} - \lambda & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} - \lambda & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} - \lambda \end{vmatrix} = 0$$

Expanding the determinant results in the following cubic equation:

$$\begin{aligned} 0 &= (\sigma_{xx} - \lambda)[(\sigma_{yy} - \lambda)(\sigma_{zz} - \lambda) - \sigma_{yz}^2] - \sigma_{xy}[\sigma_{xy}(\sigma_{zz} - \lambda) - \sigma_{yz}\sigma_{xz}] + \sigma_{xz}[\sigma_{xy}\sigma_{yz} - (\sigma_{yy} - \lambda)\sigma_{xz}] \\ &= (\sigma_{xx} - \lambda)[\sigma_{yy}\sigma_{zz} - \lambda(\sigma_{yy} + \sigma_{zz}) + \lambda^2 - \sigma_{yz}^2] - \sigma_{xy}[\sigma_{xy}(\sigma_{zz} - \lambda) - \sigma_{yz}\sigma_{xz}] + \sigma_{xz}[\sigma_{xy}\sigma_{yz} - (\sigma_{yy} - \lambda)\sigma_{xz}] \\ &= -\lambda^3 + (\sigma_{xx} + \sigma_{yy} + \sigma_{zz})\lambda^2 + (-\sigma_{yy}\sigma_{zz} - \sigma_{xx}\sigma_{yy} - \sigma_{xx}\sigma_{zz} + \sigma_{yz}^2 + \sigma_{xy}^2 + \sigma_{xz}^2)\lambda + det(\boldsymbol{\sigma}) \end{aligned}$$

---

[1] https://en.wikipedia.org/wiki/List_of_trigonometric_identities

or, after multiplying the last line by -1,

$$\lambda^3 - \mathcal{K}_1(\boldsymbol{\sigma})\lambda^2 + \mathcal{K}_2(\boldsymbol{\sigma})\lambda - \mathcal{K}_3(\boldsymbol{\sigma}) = 0 \tag{4.30}$$

with[2]:

$$
\begin{aligned}
\mathcal{K}_1(\boldsymbol{\sigma}) &= \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\
\mathcal{K}_2(\boldsymbol{\sigma}) &= \sigma_{xx}\sigma_{yy} + \sigma_{yy}\sigma_{zz} + \sigma_{xx}\sigma_{zz} - \sigma_{xy}^2 - \sigma_{xz}^2 - \sigma_{yz}^2 \\
\mathcal{K}_3(\boldsymbol{\sigma}) &= \det(\boldsymbol{\sigma}) \\
&= \sigma_{xx}\sigma_{yy}\sigma_{zz} - \sigma_{xx}\sigma_{yz}^2 - \sigma_{xy}^2\sigma_{zz} + \sigma_{xy}\sigma_{yz}\sigma_{xz} + \sigma_{xz}\sigma_{xy}\sigma_{yz} - \sigma_{xz}^2\sigma_{yy} \\
&= \sigma_{xx}\sigma_{yy}\sigma_{zz} + 2\sigma_{xy}\sigma_{yz}\sigma_{xz} - (\sigma_{xx}\sigma_{yz}^2 + \sigma_{zz}\sigma_{xy}^2 + \sigma_{yy}\sigma_{xz}^2)
\end{aligned}
\tag{4.31}
$$

$\mathcal{K}_1$, $\mathcal{K}_2$ and $\mathcal{K}_3$ are called **principal** invariants[3] (see also Appendix A.1 of Zienkiewicz & Taylor [1541] or Eq. (6.4) of Freudenthal & Geiringer [444]). These invariants can be written in a coordinate-free manner[4]:

$$
\begin{aligned}
\mathcal{K}_1(\boldsymbol{\sigma}) &= \operatorname{tr}(\boldsymbol{\sigma}) \\
\mathcal{K}_2(\boldsymbol{\sigma}) &= \frac{1}{2}(\operatorname{tr}(\boldsymbol{\sigma})^2 - \operatorname{tr}(\boldsymbol{\sigma}^2)) \\
\mathcal{K}_3(\boldsymbol{\sigma}) &= \det(\boldsymbol{\sigma})
\end{aligned}
$$

and if the stress tensor is diagonal, we have

$$
\begin{aligned}
\mathcal{K}_1(\boldsymbol{\sigma}) &= \sigma_1 + \sigma_2 + \sigma_3 \\
\mathcal{K}_2(\boldsymbol{\sigma}) &= \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3 \\
\mathcal{K}_3(\boldsymbol{\sigma}) &= \sigma_1\sigma_2\sigma_3
\end{aligned}
$$

The principal invariants $\mathcal{K}_{\{1,2,3\}}$ are related to the **moment** invariants $\mathcal{I}_{\{1,2,3\}}$ (see Section 4.3) as follows (Appendix A.2 of Zienkiewicz & Taylor [1541]):

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \mathcal{K}_1(\boldsymbol{\sigma}) \tag{4.32}$$

$$\mathcal{I}_2(\boldsymbol{\sigma}) = \frac{1}{2}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma}) \tag{4.33}$$

$$\mathcal{I}_3(\boldsymbol{\sigma}) = \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma}) \tag{4.34}$$

> write proofs in appendix

Very often we will find ourselves interested in the principal components of the deviatoric stress tensor $\boldsymbol{\tau}$ so that we now have the following determinant to compute:

$$
\begin{vmatrix}
\tau_{xx} - \lambda & \tau_{xy} & \tau_{xz} \\
\tau_{xy} & \tau_{yy} - \lambda & \tau_{yz} \\
\tau_{xz} & \tau_{yz} & \tau_{zz} - \lambda
\end{vmatrix} = 0
$$

and therefore obtain the following cubic equation

$$\lambda^3 - \mathcal{K}_1(\boldsymbol{\tau})\lambda^2 + \mathcal{K}_2(\boldsymbol{\tau})\lambda - \mathcal{K}_3(\boldsymbol{\tau}) = 0 \tag{4.35}$$

---

[2]Note that in the equation (4.30) there is often a plus sign in front of $\mathcal{K}_2$ but not always. Be careful when reading literature!

[3]https://en.wikipedia.org/wiki/Invariants_of_tensors

[4]Proofs are in Appendix Q

By definition of a deviatoric tensor we have $\mathcal{K}_1(\boldsymbol{\tau}) = 0$ and then Eqs. (4.33) and (4.34) become

$$\mathcal{I}_2(\boldsymbol{\tau}) = -\mathcal{K}_2(\boldsymbol{\tau}) \tag{4.36}$$

$$\mathcal{I}_3(\boldsymbol{\tau}) = \mathcal{K}_3(\boldsymbol{\tau}) \tag{4.37}$$

so that the cubic equation becomes

$$\lambda^3 - \mathcal{I}_2(\boldsymbol{\tau})\lambda - \mathcal{I}_3(\boldsymbol{\tau}) = 0 \tag{4.38}$$

Noting the trigonometric identity[5]

$$\sin 3\theta = 3\sin\theta - 4\sin^3\theta \qquad \text{or,} \qquad \sin^3\theta - \frac{3}{4}\sin\theta + \frac{1}{4}\sin 3\theta = 0 \tag{4.39}$$

and substituting $\lambda = r\sin\theta$ into (4.38) we have[6]

$$\sin^3\theta - \frac{\mathcal{I}_2(\boldsymbol{\tau})}{r^2}\sin\theta - \frac{\mathcal{I}_3(\boldsymbol{\tau})}{r^3} = 0 \tag{4.40}$$

Comparing (4.39) and (4.40) gives

$$r = \frac{2}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{4.41}$$

$$\sin 3\theta = -\frac{4\mathcal{I}_3(\boldsymbol{\tau})}{r^3} = -\frac{3\sqrt{3}}{2}\frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \tag{4.42}$$

The so-called Lode angle [1533] is then given by

$$\boxed{\theta_{\mathrm{L}} = \frac{1}{3}\sin^{-1}\left(-\frac{3\sqrt{3}}{2}\frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}\right)} \tag{4.43}$$

with $-\pi/6 < \theta_{\mathrm{L}} < \pi/6$. The very same equation is also found in Willett (1992) [1459] for instance.

The first root of (4.42) with θ_{L} determined for $3\theta_{\mathrm{L}}$ in the range $\pm\pi/2$ is a convenient alternative to the third invariant, $\mathcal{I}_3(\boldsymbol{\tau})$. By noting the cyclic nature of $\sin(3\theta_{\mathrm{L}} + 2n\pi)$ we have immediatly the three (and only three) possible values of $\sin\theta_{\mathrm{L}}$ which define the three principal stresses. The deviatoric principal stresses are given by $\lambda = r\sin\theta_{\mathrm{L}}$ on substitution of the three values of $\sin\theta_{\mathrm{L}}$ in turn.

We then obtain

$$\begin{Bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{Bmatrix} = \frac{2}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \begin{Bmatrix} \sin(\theta_{\mathrm{L}} + 2\pi/3) \\ \sin\theta_{\mathrm{L}} \\ \sin(\theta_{\mathrm{L}} + 4\pi/3) \end{Bmatrix} \tag{4.44}$$

with $\tau_1 > \tau_2 > \tau_3$ and $-\pi/6 \leq \theta_{\mathrm{L}} \leq \pi/6$. It is indeed easy to verify that for $-\pi/6 \leq \theta_{\mathrm{L}} \leq \pi/6$ we have $\sin(\theta_{\mathrm{L}} + 2\pi/3) > \sin\theta_{\mathrm{L}} > \sin(\theta_{\mathrm{L}} + 4\pi/3)$.

Finally, we wish to compute the principal stresses of the full stress tensor $\boldsymbol{\sigma}$. In the right coordinate system both stress and deviatoric stress tensors are diagonal and $\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau}$ writes:

$$\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} = \begin{pmatrix} -p & 0 & 0 \\ 0 & -p & 0 \\ 0 & 0 & -p \end{pmatrix} + \begin{pmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{pmatrix}$$

---

[5]see section 7.4 of Owen & Hinton [1026]

[6]Note that r and θ have nothing to do with polar, cylindrical or spherical coordinates.

so that (since $p = -\frac{1}{3}tr(\boldsymbol{\sigma}) = -\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})$)

$$\sigma_1 = \tau_1 - p = \tau_1 + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) \tag{4.45}$$

$$\sigma_2 = \tau_2 - p = \tau_2 + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) \tag{4.46}$$

$$\sigma_3 = \tau_3 - p = \tau_3 + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) \tag{4.47}$$

and finally the total principal stresses are

$$\left\{\begin{array}{c} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{array}\right\} = \frac{2}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left\{\begin{array}{c} \sin(\theta_{\mathrm{L}} + 2\pi/3) \\ \sin\theta_{\mathrm{L}} \\ \sin(\theta_{\mathrm{L}} + 4\pi/3) \end{array}\right\} + \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3}\left\{\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right\} \tag{4.48}$$

with $\sigma_1 > \sigma_2 > \sigma_3$ and $-\pi/6 \le \theta_{\mathrm{L}} \le \pi/6$. We have

$$\sin(\theta_{\mathrm{L}} + 2\pi/3) = \sin\theta_{\mathrm{L}}\cos 2\pi/3 + \cos\theta_{\mathrm{L}}\sin 2\pi/3$$
$$= -\frac{1}{2}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\frac{\sqrt{3}}{2} \tag{4.49}$$
$$\sin(\theta_{\mathrm{L}} + 4\pi/3) = \sin\theta_{\mathrm{L}}\cos 4\pi/3 + \cos\theta_{\mathrm{L}}\sin 4\pi/3$$
$$= -\frac{1}{2}\sin\theta_{\mathrm{L}} - \cos\theta_{\mathrm{L}}\frac{\sqrt{3}}{2} \tag{4.50}$$

so that

$$\left\{\begin{array}{c} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{array}\right\} = \frac{2}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left\{\begin{array}{c} -\frac{1}{2}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\frac{\sqrt{3}}{2} \\ \sin\theta_{\mathrm{L}} \\ -\frac{1}{2}\sin\theta_{\mathrm{L}} - \cos\theta_{\mathrm{L}}\frac{\sqrt{3}}{2} \end{array}\right\} + \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3}\left\{\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right\} \tag{4.51}$$

$$= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left\{\begin{array}{c} -\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}} \\ \frac{2}{\sqrt{3}}\sin\theta_{\mathrm{L}} \\ -\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} - \cos\theta_{\mathrm{L}} \end{array}\right\} + \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3}\left\{\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right\} \tag{4.52}$$

Remark. *The Lode angle is one of the Lode coordinates[7], or Haigh-Westergaard coordinates.*

Remark. *The Lode angle θ_{L} is essentially similar to the Lode parameter defined by $-\sqrt{3}\tan\theta$ [1026].*

Remark. *There are 3 different Lode angles, as explained online[8]:*

$$\sin 3\theta_s = -\sin 3\bar{\theta}_s = \cos 3\theta_c = \frac{3\sqrt{3}}{2}\frac{\mathcal{I}_3(\boldsymbol{\tau})}{(\mathcal{I}_2(\boldsymbol{\tau}))^{3/2}}$$

and they are related by $\theta_s = \frac{\pi}{6} - \theta_c$ and $\theta_s = -\bar{\theta}_s$. The one used in this document is in fact the $\bar{\theta}_s$ above.

To recap:

---

[7]https://en.wikipedia.org/wiki/Lode_coordinates
[8]https://en.wikipedia.org/wiki/Lode_coordinates

$$\sigma_1 = \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\right) \tag{4.53}$$

$$\sigma_2 = \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\frac{2}{\sqrt{3}}\sin\theta_{\mathrm{L}}\right) \tag{4.54}$$

$$\sigma_3 = \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} - \cos\theta_{\mathrm{L}}\right) \tag{4.55}$$

We will later need $\sigma_1 - \sigma_3$ and $\sigma_1 + \sigma_3$ so we compute these quantities hereafter:

$$\begin{aligned}
\sigma_1 - \sigma_3 &= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}} + \frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\right) \\
&= 2\cos\theta_{\mathrm{L}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{4.56} \\
\sigma_1 + \sigma_3 &= \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\right) + \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} - \cos\theta_{\mathrm{L}}\right) \\
&= \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma}) - \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\frac{2}{\sqrt{3}}\sin\theta_{\mathrm{L}} \tag{4.57}
\end{aligned}$$

or,

$$\frac{\sigma_1 - \sigma_3}{2} = \cos\theta_{\mathrm{L}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{4.58}$$

$$\frac{\sigma_1 + \sigma_3}{2} = \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) - \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}} \tag{4.59}$$

Remark. *The expression for the Lode angle is different in [1543, p101] than in [1533] or [1541, p62]. They all look suspiciously wrong too.*

4.2.3   About the 2nd principal invariant of the deviatoric stress

$$\begin{aligned}
\mathcal{K}_2(\boldsymbol{\tau}) &= \frac{1}{2}[\mathrm{Tr}(\boldsymbol{\tau})^2 - \mathrm{Tr}(\boldsymbol{\tau}^2)] \\
&= \frac{1}{2}[(\tau_{xx} + \tau_{yy})^2 - (\tau_{xx}^2 + 2\tau_{xy}^2 + \tau_{yy}^2)] \\
&= \frac{1}{2}[\tau_{xx}^2 + 2\tau_{xx}\tau_{yy} + \tau_{yy}^2 - \tau_{xx}^2 - 2\tau_{xy}^2 - \tau_{yy}^2] \\
&= \frac{1}{2}[2\tau_{xx}\tau_{yy} - 2\tau_{xy}^2] \\
&= \tau_{xx}\tau_{yy} - \tau_{xy}^2 \\
&= \left(\sigma_{xx} - \frac{\sigma_{xx} + \sigma_{yy}}{2}\right)\left(\sigma_{yy} - \frac{\sigma_{xx} + \sigma_{yy}}{2}\right) - \tau_{xy}^2 \\
&= \left(\sigma_{xx} - \frac{\sigma_{xx} + \sigma_{yy}}{2}\right)\left(\sigma_{yy} - \frac{\sigma_{xx} + \sigma_{yy}}{2}\right) - \tau_{xy}^2 \\
&= \left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)\left(\frac{-\sigma_{xx} + \sigma_{yy}}{2}\right) - \tau_{xy}^2 \\
&= -\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 - \tau_{xy}^2
\end{aligned}$$

Looking at Eq. (4.29), we can then write

$$\tau_{max} = \sqrt{-\mathcal{K}_2(\boldsymbol{\tau})} = \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2}$$

4.3   Tensor (moment) invariants

There are many different notations used in the literature for invariants and these can prove to be confusing[9]. Note that we only consider symmetric tensors in what follows. Given a tensor $\boldsymbol{T}$, one can compute its (moment) invariants as follows (see [1118, p.339], or Appendix A.2 of [1541])

$$\begin{aligned}
\mathcal{I}_1(\boldsymbol{T}) &= \text{tr}[\boldsymbol{T}] & (4.60)\\
&= T_{xx} + T_{yy} + T_{zz} & (4.61)\\
\mathcal{I}_2(\boldsymbol{T}) &= \frac{1}{2}\text{tr}[\boldsymbol{T} \cdot \boldsymbol{T}] & (4.62)\\
&= \frac{1}{2}\sum_{ij} T_{ij}T_{ji} & (4.63)\\
&= \frac{1}{2}(T_{xx}^2 + T_{yy}^2 + T_{zz}^2) + T_{xy}^2 + T_{xz}^2 + T_{yz}^2 & (4.64)\\
\mathcal{I}_3(\boldsymbol{T}) &= \frac{1}{3}\text{tr}[\boldsymbol{T} \cdot \boldsymbol{T} \cdot \boldsymbol{T}] & (4.65)\\
&= \frac{1}{3}\sum_i \sum_j \sum_k T_{ij}T_{jk}T_{ki} & (4.66)
\end{aligned}$$

---

[9]No kidding, true story.

i	j	k	$T_{ij}T_{jk}T_{ki}$	$symm$
x	x	x	$T_{xx}T_{xx}T_{xx}$	T_{xx}^3
y	x	x	$T_{yx}T_{xx}T_{xy}$	$T_{xx}T_{xy}^2$
z	x	x	$T_{zx}T_{xx}T_{xz}$	$T_{xx}T_{xz}^2$
x	y	x	$T_{xy}T_{yx}T_{xx}$	$T_{xx}T_{xy}^2$
y	y	x	$T_{yy}T_{yx}T_{xy}$	$T_{yy}T_{xy}^2$
z	y	x	$T_{zy}T_{yx}T_{xz}$	$T_{xy}T_{xz}T_{yz}$
x	z	x	$T_{xz}T_{zx}T_{xx}$	$T_{xx}T_{xz}^2$
y	z	x	$T_{yz}T_{zx}T_{xy}$	$T_{xy}T_{xz}T_{yz}$
z	z	x	$T_{zz}T_{zx}T_{xz}$	$T_{zz}T_{xz}^2$
x	x	y	$T_{xx}T_{xy}T_{yx}$	$T_{xx}T_{xy}^2$
y	x	y	$T_{yx}T_{xy}T_{yy}$	$T_{yy}T_{xy}^2$
z	x	y	$T_{zx}T_{xy}T_{yz}$	$T_{xy}T_{xz}T_{yz}$
x	y	y	$T_{xy}T_{yy}T_{yx}$	$T_{yy}T_{xy}^2$
y	y	y	$T_{yy}T_{yy}T_{yy}$	T_{yy}^3
z	y	y	$T_{zy}T_{yy}T_{yz}$	$T_{yy}T_{yz}^2$
x	z	y	$T_{xz}T_{zy}T_{yx}$	$T_{xy}T_{xz}T_{yz}$
y	z	y	$T_{yz}T_{zy}T_{yy}$	$T_{yy}T_{yz}^2$
z	z	y	$T_{zz}T_{zy}T_{yz}$	$T_{zz}T_{yz}^2$
x	x	z	$T_{xx}T_{xz}T_{zx}$	$T_{xx}T_{xz}^2$
y	x	z	$T_{yx}T_{xz}T_{zy}$	$T_{xy}T_{xz}T_{yz}$
z	x	z	$T_{zx}T_{xz}T_{zz}$	$T_{zz}T_{xz}^2$
x	y	z	$T_{xy}T_{yz}T_{zx}$	$T_{xy}T_{yz}T_{yz}$
y	y	z	$T_{yy}T_{yz}T_{zy}$	$T_{yy}T_{yz}^2$
z	y	z	$T_{zy}T_{yz}T_{zz}$	$T_{zz}T_{yz}^2$
x	z	z	$T_{xz}T_{zz}T_{zx}$	$T_{zz}T_{xz}^2$
y	z	z	$T_{yz}T_{zz}T_{zy}$	$T_{zz}T_{yz}^2$
z	z	z	$T_{zz}T_{zz}T_{zz}$	T_{zz}^3

In the end

$$\sum_{i=x,y,z}\sum_{j=x,y,z}\sum_{k=x,y,z} T_{ij}T_{jk}T_{ki} = T_{xx}(T_{xx}^2+3T_{xy}^2+3T_{xz}^2)+T_{yy}(3T_{xy}^2+T_{yy}^2+3T_{yz}^2)+T_{zz}(3T_{xz}^2+3T_{yz}^2+T_{zz}^2)+6T_{xy}T_{yz}T_{yz}$$

and then the third moment invariant of the symmetric tensor $\boldsymbol{T}$ is given by:

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{T}) &= \frac{1}{3}T_{xx}(T_{xx}^2 + 3T_{xy}^2 + 3T_{xz}^2) \\
&+ \frac{1}{3}T_{yy}(3T_{xy}^2 + T_{yy}^2 + 3T_{yz}^2) \\
&+ \frac{1}{3}T_{zz}(3T_{xz}^2 + 3T_{yz}^2 + T_{zz}^2) \\
&+ 2T_{xy}T_{xz}T_{yz} \\
&= \frac{1}{3}(T_{xx}^3 + T_{yy}^3 + T_{zz}^3) + T_{xx}(T_{xy}^2 + T_{xz}^2) + T_{yy}(T_{xy}^2 + T_{yz}^2) + T_{zz}(T_{xz}^2 + T_{yz}^2) + 2T_{xy}T_{xz}T_{yz}
\end{aligned}
$$

4.4 Stress & strain rate (moment) invariants

The implementation of the plasticity criterions relies essentially on the (moment) invariants $\mathcal{I}_{1,2,3}$ of the (deviatoric) stress $\boldsymbol{\tau}$ and the (deviatoric) strainrate tensors $\dot{\boldsymbol{\varepsilon}}$:

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \tag{4.67}$$

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2) + \tau_{xy}^2 + \tau_{xz}^2 + \tau_{yz}^2 \tag{4.68}$$

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{1}{3}\tau_{xx}(\tau_{xx}^2 + 3\tau_{xy}^2 + 3\tau_{xz}^2)$$

$$+ \frac{1}{3}\tau_{yy}(3\tau_{xy}^2 + \tau_{yy}^2 + 3\tau_{yz}^2)$$

$$+ \frac{1}{3}\tau_{zz}(3\tau_{xz}^2 + 3\tau_{yz}^2 + \tau_{zz}^2)$$

$$+ 2\tau_{xy}\tau_{xz}\tau_{yz} \tag{4.69}$$

and also the second invariant of the deviatoric strain rate is:

$$\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d) = \frac{1}{2}\left[(\dot{\varepsilon}_{xx}^d)^2 + (\dot{\varepsilon}_{yy}^d)^2 + (\dot{\varepsilon}_{zz}^d)^2\right] + (\dot{\varepsilon}_{xy}^d)^2 + (\dot{\varepsilon}_{xz}^d)^2 + (\dot{\varepsilon}_{yz}^d)^2$$

$$= \frac{1}{6}\left[(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{zz})^2 + (\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{zz})^2\right] + \dot{\varepsilon}_{xy}^2 + \dot{\varepsilon}_{xz}^2 + \dot{\varepsilon}_{yz}^2 \tag{4.70}$$

Proofs of these relationships are given in Appendix Q.

We have

$$\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2 = \left(\sigma_{xx} - \frac{1}{3}I_1\right)^2 + \left(\sigma_{yy} - \frac{1}{3}I_1\right)^2 + \left(\sigma_{zz} - \frac{1}{3}I_1\right)^2$$

$$= \sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{2}{3}I_1(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) + 3\frac{1}{9}I_1^2$$

$$= \sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{2}{3}I_1^2 + \frac{1}{3}I_1^2$$

$$= \sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{1}{3}I_1^2$$

$$= \sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{1}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})^2$$

$$= \sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{1}{3}(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xx}\sigma_{yy} + 2\sigma_{xx}\sigma_{zz} + 2\sigma_{yy}\sigma_{zz})$$

$$= \frac{1}{3}(3\sigma_{xx}^2 + 3\sigma_{yy}^2 + 3\sigma_{zz}^2 - \sigma_{xx}^2 - \sigma_{yy}^2 - \sigma_{zz}^2 - 2\sigma_{xx}\sigma_{yy} - 2\sigma_{xx}\sigma_{zz} - 2\sigma_{yy}\sigma_{zz})$$

$$= \frac{1}{3}(2\sigma_{xx}^2 + 2\sigma_{yy}^2 + 2\sigma_{zz}^2 - 2\sigma_{xx}\sigma_{yy} - 2\sigma_{xx}\sigma_{zz} - 2\sigma_{yy}\sigma_{zz}) \tag{4.71}$$

$$= \frac{1}{3}((\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{xx} - \sigma_{zz})^2 + (\sigma_{yy} - \sigma_{zz})^2) \tag{4.72}$$

so that

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6}\left[(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{xx} - \sigma_{zz})^2\right] + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2$$

Remark. *$\mathcal{I}_2(\boldsymbol{\tau})$ is often called J_2 or J_2' so that one sometimes speaks of J_2-plasticity.*

These (second) invariants are almost always used under a square root so we define:

$$\tau_e = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \qquad \dot{\varepsilon}_e = \sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d)} \tag{4.73}$$

Note that these quantities have the same dimensions as their tensor counterparts, i.e. Pa for stresses and s^{-1} for strain rates.

If the stress tensor is such that it is diagonal, i.e.

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\tau} = \begin{pmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{pmatrix}$$

then the invariants are

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_1 + \sigma_2 + \sigma_3 \tag{4.74}$$

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6} \left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2 \right] \tag{4.75}$$

$$\mathcal{I}_3(\boldsymbol{\tau}) = \tau_1 \tau_2 \tau_3$$

$$= \frac{1}{3} \text{tr}[\boldsymbol{\tau} \cdot \boldsymbol{\tau} \cdot \boldsymbol{\tau}]$$

$$= \frac{1}{3} \text{tr} \left[\begin{pmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{pmatrix} \cdot \begin{pmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{pmatrix} \cdot \begin{pmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{pmatrix} \right]$$

$$= \frac{1}{3} \text{tr} \begin{pmatrix} \tau_1^3 & 0 & 0 \\ 0 & \tau_2^3 & 0 \\ 0 & 0 & \tau_3^3 \end{pmatrix}$$

$$= \frac{1}{3} (\tau_1^3 + \tau_2^3 + \tau_3^3)$$

$$= \frac{1}{3} [(\sigma_1 - \mathcal{I}_1(\boldsymbol{\sigma})/3)^3 + (\sigma_2 - \mathcal{I}_1(\boldsymbol{\sigma})/3)^3 + (\sigma_3 - \mathcal{I}_1(\boldsymbol{\sigma})/3)^3]$$

$$= \frac{1}{3 \cdot 27} [(3\sigma_1 - \mathcal{I}_1(\boldsymbol{\sigma}))^3 + (3\sigma_2 - \mathcal{I}_1(\boldsymbol{\sigma}))^3 + (3\sigma_3 - \mathcal{I}_1(\boldsymbol{\sigma}))^3]$$

$$= \frac{1}{81} \left[(2\sigma_1 - \sigma_2 - \sigma_3)^3 + (2\sigma_2 - \sigma_1 - \sigma_3)^3 + (2\sigma_3 - \sigma_1 - \sigma_2)^3 \right] \tag{4.76}$$

The formulation of the third invariant of $\boldsymbol{\tau}$ in Eq. 4.76 is used in Wojciechowski [1467].

4.5 Two-dimensional plane strain calculations

We start from the 3D strain rate tensor

$$\dot{\boldsymbol{\varepsilon}}(\vec{v}) = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} & \dot{\varepsilon}_{yz} \\ \dot{\varepsilon}_{zx} & \dot{\varepsilon}_{zy} & \dot{\varepsilon}_{zz} \end{pmatrix}$$

The plane strain assumption is such that the problem at hand is assumed to be infinite in a given direction. In the case of computational geodynamics, most 2D modelling is a vertical section of the crust-lithosphere-mantle and the underlying implicit assumption is then that the orogen/rift/subduction/etc ... is infinite in the direction perpendicular to the screen/paper.

Let us assume that the deformation takes place in the x,y-plane, so that $w = 0$ (velocity in the z direction is zero) and $\partial_z \to 0$ (no change in the z direction). We then have $\dot{\varepsilon}_{zz} = 0$ as well as $\dot{\varepsilon}_{xz} = 0$ and $\dot{\varepsilon}_{yz} = 0$, so that the strain rate tensor is

$$\dot{\boldsymbol{\varepsilon}}(\vec{v}) = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & 0 \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Incompressible flow

If the flow is incompressible then the deviatoric stress tensor is given by

$$\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) = 2\eta\left(\dot{\boldsymbol{\varepsilon}}(\vec{v}) - \frac{1}{3}\underbrace{\text{tr}[\dot{\boldsymbol{\varepsilon}}]}_{=0}\mathbf{1}\right) = 2\eta\dot{\boldsymbol{\varepsilon}}(\vec{v}) = \begin{pmatrix} \tau_{xx} & \tau_{xy} & 0 \\ \tau_{yx} & \tau_{yy} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

One then discards the unnecessary line and column in the tensor, leaving a 2×2 matrix. Finding the principal stress components is then trivial since we have done it in 2D already.

It is important to keep in mind that the invariants we need to implement the rheologies are $\mathcal{I}_1(\boldsymbol{\sigma})$, $\mathcal{I}_2(\boldsymbol{\tau})$ and $\mathcal{I}_3(\boldsymbol{\tau})$. By formulating our yield surfaces with pressure $p = -\mathcal{I}_1(\boldsymbol{\sigma})/3$ we can then avoid confusion, and since the other two invariants are functions of $\boldsymbol{\tau}$ the pressure term does not pose any problem: simply set τ_{xz}, τ_{yz} and τ_{zz} to zero in the equations of Section 4.4 and we obtain:

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2) + \tau_{xy}^2 \tag{4.77}$$

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{1}{3}\tau_{xx}(\tau_{xx}^2 + 3\tau_{xy}^2) + \frac{1}{3}\tau_{yy}(3\tau_{xy}^2 + \tau_{yy}^2)$$

$$= \frac{1}{3}(\tau_{xx}^3 + 3\tau_{xx}\tau_{xy}^2 + 3\tau_{yy}\tau_{xy}^2 + \tau_{yy}^3)$$

$$= \frac{1}{3}(\tau_{xx}^3 + 3(\tau_{xx} + \tau_{yy})\tau_{xy}^2 + \tau_{yy}^3)$$

$$= \frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^3) \qquad \text{since } \tau_{ii} = 0 \tag{4.78}$$

The principal stresses of the deviatoric stress tensor $\boldsymbol{\tau}$ are given by

$$\tau_1 = \frac{\tau_{xx} + \tau_{yy}}{2} + \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2}$$

$$\tau_2 = \frac{\tau_{xx} + \tau_{yy}}{2} - \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2} \tag{4.79}$$

The full stress tensor is then

$$\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau} = \begin{pmatrix} -p + \tau_{xx} & \tau_{xy} & 0 \\ \tau_{yx} & -p + \tau_{yy} & 0 \\ 0 & 0 & -p \end{pmatrix}$$

so it remains a 3×3 tensor!

However, looking at the conservation of momentum,

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \rho\vec{g} = \vec{0}$$

Given the conditions for plane-strain then $\vec{g}$ is likely to be in the xy-plane so that the z component of the equation becomes:

$$-\partial_z p = 0$$

and since we have $\partial_z \to 0$ anyways this equation is automatically fulfilled. Then, we might as well proceed by considering that the stress tensor is in fact 2D as the third row/column has no incidence. In that case the pressure is given by $p = -\mathcal{I}_1(\boldsymbol{\sigma})/2$. In the plasticity yield criterion or plastic potential we will need the full stress $\boldsymbol{\sigma}$ only via its first invariant (i.e. the pressure). The other two invariants are those of the deviatoric stress.

Let us start from the deviatoric stress tensor:

$$\boldsymbol{\tau} = \boldsymbol{\sigma} - \frac{1}{2}\mathcal{I}_1(\boldsymbol{\sigma}) = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} - \frac{\sigma_{xx} + \sigma_{yy}}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} (\sigma_{xx} - \sigma_{yy})/2 & \sigma_{xy} \\ \sigma_{xy} & -(\sigma_{xx} - \sigma_{yy})/2 \end{pmatrix}$$

The second invariant of the deviatoric stress tensor is then

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2}\left(2(\sigma_{xx} - \sigma_{yy})^2/4 + 2\sigma_{xy}^2\right)$$

or

$$\mathcal{I}_2(\boldsymbol{\tau}) = \left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2$$

and the effective deviatoric stress

$$\tau_e = \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2}$$

Remark: Using the form of $\mathcal{I}_2(\boldsymbol{\tau})$ above one arrives at

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xx}} = 2\frac{1}{2}\frac{\sigma_{xx} - \sigma_{yy}}{2} = \tau_{xx}$$

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yy}} = -2\frac{1}{2}\frac{\sigma_{xx} - \sigma_{yy}}{2} = \tau_{yy}$$

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xy}} = 2\sigma_{xy} = 2\tau_{xy}$$

which is ...wrong! One should first write the second invariant for the generic case of the deviatoric stress tensor (without assuming it is symmetric):

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2}\left(2(\sigma_{xx} - \sigma_{yy})^2/4 + \sigma_{xy}^2 + \sigma_{yx}^2\right) = \left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \frac{1}{2}\sigma_{xy}^2 + \frac{1}{2}\sigma_{yx}^2$$

Then

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xx}} = 2\frac{1}{2}\frac{\sigma_{xx} - \sigma_{yy}}{2} = \tau_{xx}$$

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yy}} = -2\frac{1}{2}\frac{\sigma_{xx} - \sigma_{yy}}{2} = \tau_{yy}$$

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xy}} = \sigma_{xy} = \tau_{xy}$$

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yx}} = \sigma_{yx} = \tau_{yx}$$

which can be simply written as

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} = \boldsymbol{\tau}$$

Compressible flow

If the flow is not incompressible, then the deviatoric strain rate tensor is

$$\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}) = \dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}}) - \frac{1}{3}\mathrm{tr}[\dot{\boldsymbol{\varepsilon}}]\mathbf{1} = \dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}}) - \frac{1}{3}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy})\mathbf{1} = \begin{pmatrix} \frac{2}{3}\dot{\varepsilon}_{xx} - \frac{1}{3}\dot{\varepsilon}_{yy} & \dot{\varepsilon}_{xy} & 0 \\ \dot{\varepsilon}_{yx} & -\frac{1}{3}\dot{\varepsilon}_{xx} + \frac{2}{3}\dot{\varepsilon}_{yy} & 0 \\ 0 & 0 & -\frac{1}{3}\dot{\varepsilon}_{xx} - \frac{1}{3}\dot{\varepsilon}_{yy} \end{pmatrix}$$

The deviatoric stress tensor now has the form

$$\boldsymbol{\tau} = \begin{pmatrix} \tau_{xx} & \tau_{xy} & 0 \\ \tau_{yx} & \tau_{yy} & 0 \\ 0 & 0 & \tau_{zz} \end{pmatrix}$$

We are interested in the principal components of the deviatoric stress tensor $\boldsymbol{\tau}$ so that we now have the following determinant to compute:

$$\begin{vmatrix} \tau_{xx} - \lambda & \tau_{xy} & 0 \\ \tau_{xy} & \tau_{yy} - \lambda & 0 \\ 0 & 0 & \tau_{zz} - \lambda \end{vmatrix} = 0$$

which yields the following characteristic equation:

$$(\tau_{zz} - \lambda)(\lambda - \tau_1)(\lambda - \tau_2) = 0$$

where $\tau_{1,2}$ have previously been obtained in the 2D case:

$$\begin{aligned} \tau_1 &= \frac{\tau_{xx} + \tau_{yy}}{2} + \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2} \\ \tau_2 &= \frac{\tau_{xx} + \tau_{yy}}{2} - \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2} \end{aligned} \tag{4.80}$$

We have

$$\begin{aligned} \tau_{xx} + \tau_{yy} &= 2\eta\frac{1}{3}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy}) \\ \tau_{xx} - \tau_{yy} &= 2\eta(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy}) \end{aligned} \tag{4.81}$$

Then

$$\begin{aligned} \tau_1 &= \frac{\tau_{xx} + \tau_{yy}}{2} + \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2} \\ &= \eta\frac{1}{3}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy}) + \eta\sqrt{(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + 4\dot{\varepsilon}_{xy}^2} \\ \tau_2 &= \frac{\tau_{xx} + \tau_{yy}}{2} - \sqrt{\left(\frac{\tau_{xx} - \tau_{yy}}{2}\right)^2 + \tau_{xy}^2} \\ &= \eta\frac{1}{3}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy}) - \eta\sqrt{(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + 4\dot{\varepsilon}_{xy}^2} \end{aligned} \tag{4.82}$$

It does not look like it is going to simplify down the road ... Also, the third eigenvalue/principal stress remains and it is not clear whether it is larger or smaller than the other two. The 3D framework is then probably the most appropriate.

Let us now turn to the second invariant of the deviatoric strain rate (see Eq. (4.70)):

$$
\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d) = \frac{1}{2}\dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d \tag{4.83}
$$

$$
= \frac{1}{2}\left[(\dot{\varepsilon}_{xx}^d)^2 + (\dot{\varepsilon}_{yy}^d)^2 + (\dot{\varepsilon}_{zz}^d)^2\right] + (\dot{\varepsilon}_{xy}^d)^2 + (\dot{\varepsilon}_{xz}^d)^2 + (\dot{\varepsilon}_{yz}^d)^2 \tag{4.84}
$$

But there is also an expression for $\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d)$ directly as a function of the $\dot{\varepsilon}_{ij}$ components (see Eq. (4.70)):

$$
\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d) = \frac{1}{6}\left[(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{zz})^2 + (\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{zz})^2\right] + \dot{\varepsilon}_{xy}^2 + \dot{\varepsilon}_{xz}^2 + \dot{\varepsilon}_{yz}^2 \tag{4.85}
$$

$$
= \frac{1}{6}\left[(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{xx})^2\right] + \dot{\varepsilon}_{xy}^2 \tag{4.86}
$$

$$
= \frac{1}{6}\left[\dot{\varepsilon}_{xx}^2 - 2\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + \dot{\varepsilon}_{yy}^2 + \dot{\varepsilon}_{yy}^2 + \dot{\varepsilon}_{xx}^2\right] + \dot{\varepsilon}_{xy}^2 \tag{4.87}
$$

$$
= \frac{1}{6}\left[2\dot{\varepsilon}_{xx}^2 - 2\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{yy}^2\right] + \dot{\varepsilon}_{xy}^2 \tag{4.88}
$$

$$
= \frac{1}{3}\left[\dot{\varepsilon}_{xx}^2 - \dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + \dot{\varepsilon}_{yy}^2\right] + \dot{\varepsilon}_{xy}^2 \tag{4.89}
$$

If we now do things the old/wrong(?) way, one would start directly from the 2D strain rate tensor

$$
\dot{\boldsymbol{\varepsilon}} = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} \end{pmatrix}
$$

The deviatoric strain rate tensor is then logically defined as

$$
\dot{\boldsymbol{\varepsilon}}^d = \dot{\boldsymbol{\varepsilon}} - \frac{1}{2}Tr[\dot{\boldsymbol{\varepsilon}}]\mathbf{1} = \dot{\boldsymbol{\varepsilon}} - \frac{1}{2}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy})\mathbf{1}
$$

or,

$$
\dot{\boldsymbol{\varepsilon}}^d = \begin{pmatrix} \frac{1}{2}\dot{\varepsilon}_{xx} - \frac{1}{2}\dot{\varepsilon}_{yy} & \dot{\varepsilon}_{xy} \\ \dot{\varepsilon}_{yx} & -\frac{1}{2}\dot{\varepsilon}_{xx} + \frac{1}{2}\dot{\varepsilon}_{yy} \end{pmatrix}
$$

Let us now turn to the second invariant of the deviatoric strain rate (see Section 3.21 in fieldstone)

$$
\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d) = \frac{1}{2}\dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d
$$

$$
= \frac{1}{2}[(\frac{1}{2}\dot{\varepsilon}_{xx} - \frac{1}{2}\dot{\varepsilon}_{yy})^2 + (-\frac{1}{2}\dot{\varepsilon}_{xx} + \frac{1}{2}\dot{\varepsilon}_{yy})^2] + \dot{\varepsilon}_{xy}^2
$$

$$
= \frac{1}{2}[\frac{1}{4}(2\dot{\varepsilon}_{xx}^2 - 4\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{yy}^2)] + \dot{\varepsilon}_{xy}^2
$$

$$
= \frac{1}{4}[\dot{\varepsilon}_{xx}^2 - 2\dot{\varepsilon}_{xx}\dot{\varepsilon}_{yy} + \dot{\varepsilon}_{yy}^2] + \dot{\varepsilon}_{xy}^2 \tag{4.90}
$$

which is not the same as the previous expression!

4.6   Alternative principal stresses notations

The principal stresses of the stress tensor $\boldsymbol{\sigma}$ are σ_1, σ_2 and σ_3 with $\sigma_1 \geq \sigma_2 \geq \sigma_3$. Following Wojciechowski [1467], we start by stating that the intermediate principal stress can always be represented as a linear combination of two other stresses:

$$\sigma_2 = (1 - b)\sigma_1 + b\sigma_3 \qquad \text{where} \qquad b = \frac{\sigma_1 - \sigma_2}{\sigma_1 - \sigma_3} \in [0, 1] \tag{4.91}$$

The quantity b is called the principal stress ratio. Let us now introduce the maximum shear plane stresses σ_m and τ_m such that [10]

$$\boxed{\sigma_m = \frac{\sigma_1 + \sigma_3}{2}} \qquad \boxed{\tau_m = \frac{\sigma_1 - \sigma_3}{2}} \tag{4.92}$$

so that we have

$$\sigma_1 = \sigma_m + \tau_m \tag{4.93}$$
$$\sigma_2 = \sigma_m - a\tau_m \tag{4.94}$$
$$\sigma_3 = \sigma_m - \tau_m \tag{4.95}$$

The quantity $a \in [-1, 1]$ is an equivalent measure of the principal stress ratio and is defined as

$$a = 2b - 1 = 2\frac{\sigma_1 - \sigma_2}{\sigma_1 - \sigma_3} - 1 = \frac{\sigma_1 - 2\sigma_2 + \sigma_3}{\sigma_1 - \sigma_3} \tag{4.96}$$

We can introduce a, σ_m and τ_m in the invariants above:

$$
\begin{aligned}
\mathcal{I}_1(\boldsymbol{\sigma}) &= \sigma_1 + \sigma_2 + \sigma_3 \\
&= (\sigma_m + \tau_m) + (\sigma_m - a\tau_m) + (\sigma_m - \tau_m) \\
&= 3\sigma_m - a\tau_m
\end{aligned}
\tag{4.97}
$$

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{6}\left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2\right] \\
&= \frac{1}{6}\left[(\sigma_m + \tau_m - \sigma_m + a\tau_m)^2 + (\sigma_m - a\tau_m - \sigma_m + \tau_m)^2 + (\sigma_m + \tau_m - \sigma_m + \tau_m)^2\right] \\
&= \frac{1}{6}\left[(\tau_m + a\tau_m)^2 + (-a\tau_m + \tau_m)^2 + (\tau_m + \tau_m)^2\right] \\
&= \frac{\tau_m^2}{6}\left[(1 + a)^2 + (-a + 1)^2 + 4\right] \\
&= \frac{\tau_m^2}{6}\left[1 + 2a + a^2 + 1 - 2a + a^2 + 4\right] \\
&= \frac{\tau_m^2}{3}\left(a^2 + 3\right)
\end{aligned}
\tag{4.98}
$$

[10]Although most of this section is inspired by Wojciechowski [1467], I have decided not to use his notations which are very confusing since he denotes σ_m by p

Using the definition of the third invariant of Eq. (4.76):

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{1}{81}\left[(2\sigma_1 - \sigma_2 - \sigma_3)^3 + (2\sigma_2 - \sigma_1 - \sigma_3)^3 + (2\sigma_3 - \sigma_1 - \sigma_2)^3\right] \\
&= \frac{1}{81}\left[(2\sigma_m + 2\tau_m - \sigma_m + a\tau_m - \sigma_m + \tau_m)^3 + (2\sigma_m - 2a\tau_m - \sigma_m - \tau_m - \sigma_m + \tau_m)^3 + (2\sigma_m - 2\tau_m - \right. \\
&= \frac{1}{81}\left[(2\tau_m + a\tau_m + \tau_m)^3 + (-2a\tau_m - \tau_m + \tau_m)^3 + (-2\tau_m - \tau_m + a\tau_m)^3\right] \\
&= \frac{\tau_m^3}{81}\left[(3+a)^3 + (-2a)^3 + (-3+a)^3\right] \\
&= \frac{\tau_m^3}{81}\left[27 + 9a + 3a^2 + a^3 - 8a^3 - 27 + 9a - 3a^2 + a^3\right] \\
&= \frac{\tau_m^3}{81}\left(18a - 6a^3\right) \\
&= \frac{2a\tau_m^3}{27}\left(3 - a^2\right)
\end{aligned}
$$

which is different than Eq. (14) of Wojciechowski [1467]!!

To recap:

$$
\boxed{\mathcal{I}_1(\boldsymbol{\sigma}) = 3\sigma_m - a\tau_m} \qquad \boxed{\mathcal{I}_2(\boldsymbol{\tau}) = \frac{\tau_m^2}{3}\left(a^2 + 3\right)} \qquad \boxed{\mathcal{I}_3(\boldsymbol{\tau}) = \frac{2a\tau_m^3}{27}\left(3 - a^2\right)} \qquad (4.100)
$$

Remark. *Wojciechowski [1467] defines the Lode angle as being the opposite of my definition in Eq. 4.43.*

Finally, we can show using Eqs. (4.53,4.54,4.55) that

$$
\begin{aligned}
a &= \frac{\sigma_1 - 2\sigma_2 + \sigma_3}{\sigma_1 - \sigma_3} \\
&= \frac{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta + \cos\theta\right) - 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\frac{2}{\sqrt{3}}\sin\theta\right) + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta - \cos\theta\right)}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta + \cos\theta\right) - \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\frac{1}{\sqrt{3}}\sin\theta - \cos\theta\right)} \\
&= \frac{\left(-\frac{1}{\sqrt{3}}\sin\theta + \cos\theta\right) - 2\left(\frac{2}{\sqrt{3}}\sin\theta\right) + \left(-\frac{1}{\sqrt{3}}\sin\theta - \cos\theta\right)}{\left(-\frac{1}{\sqrt{3}}\sin\theta + \cos\theta\right) - \left(-\frac{1}{\sqrt{3}}\sin\theta - \cos\theta\right)} \\
&= \frac{-\frac{6}{\sqrt{3}}\sin\theta}{2\cos\theta} \\
&= -\frac{3}{\sqrt{3}}\frac{\sin\theta}{\cos\theta} \\
&= -\sqrt{3}\tan\theta \qquad\qquad\qquad (4.101)
\end{aligned}
$$

Here again we arrive at the opposite of Eq. (16) of Wojciechowski [1467].

4.7 Recap of notations and definitions of stress invariants

When it comes to stress invariants, I urge the reader to be extremely careful when considering their source(s). As we have seen these come in two flavours (e.g. principal and moment invariants) and they are often written for the full stress or deviatoric tensor. On top of it all, typos are common like in any source and the occasional minus sign or factor 2 or 3 can be missing. This is the reason why I have spent substantial time re-deriving those in the past pages with a consistent set of notations:

$\boldsymbol{\sigma}$	(full) stress tensor
$\sigma_1,\ \sigma_2,\ \sigma_3$	principal stresses
$\boldsymbol{\tau}$	deviatoric stress tensor
$\tau_1,\ \tau_2,\ \tau_3$	principal deviatoric stresses
$\mathcal{I}_1(\boldsymbol{T})$	first moment invariant of tensor $\boldsymbol{T}$
$\mathcal{I}_2(\boldsymbol{T})$	second moment invariant of tensor $\boldsymbol{T}$
$\mathcal{I}_3(\boldsymbol{T})$	third moment invariant of tensor $\boldsymbol{T}$
$\tau_e = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$	effective deviatoric stress
$\dot{\varepsilon}_e = \sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}}^d)}$	effective deviatoric strain rate

Proofs of all the following relationships are given in Appendix Q.

- First invariant

$$
\begin{aligned}
\mathcal{I}_1(\boldsymbol{\sigma}) &= \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\
\mathcal{I}_1(\boldsymbol{\tau}) &= 0 \\
\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} &= 1
\end{aligned}
$$

- Second invariant

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\boldsymbol{\tau} : \boldsymbol{\tau} \\
&= \frac{1}{2}\mathrm{tr}[\boldsymbol{\tau} \cdot \boldsymbol{\tau}] \\
&= \frac{1}{2}\sum_{ij} \tau_{ij}\tau_{ji} \\
&= \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2 + 2\tau_{xy}^2 + 2\tau_{xz}^2 + 2\tau_{yz}^2) \\
&= \frac{1}{6}\left[(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{xx} - \sigma_{zz})^2\right] + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma}) \\
\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} &= \boldsymbol{\tau}
\end{aligned}
$$

- Third invariant

$$\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{1}{3}\sum_{ijk}\tau_{ij}\tau_{jk}\tau_{ki} \\
&= \det(\boldsymbol{\tau}) \\
&= \frac{1}{3}\mathrm{tr}[\boldsymbol{\tau}\cdot\boldsymbol{\tau}\cdot\boldsymbol{\tau}] \\
&= \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma}) + \mathcal{I}_3(\boldsymbol{\sigma})
\end{aligned}$$

$$\frac{\partial\mathcal{I}_3(\boldsymbol{\tau})}{\partial\boldsymbol{\sigma}} = \tag{4.102}$$

$$\theta_{\mathrm{L}} = \frac{1}{3}\sin^{-1}\left(-\frac{3\sqrt{3}}{2}\frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}\right) \tag{4.103}$$

$$\frac{\partial\mathcal{I}_1(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}} = \mathbf{1} \tag{4.104}$$

$$\frac{\partial\mathcal{I}_2(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}} = \boldsymbol{\sigma} \tag{4.105}$$

$$\frac{\partial\mathcal{I}_3(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}} = \boldsymbol{\sigma}\cdot\boldsymbol{\sigma} \tag{4.106}$$

Chapter 5

Rheology

The reader is referred to Barnes [49] for a discussion and review of non-linear viscous rheologies and to Coussot [300] for a review of experimental data for yield stress fluid flows. See also Tanner & Tanner [1321] for a summary of Heinrich Hencky's scientific work on rheology.

5.1 Relevant literature

- Plasticity and Geomechanics, Davis and Selvadurai. [338]

- Elasticity and Geomechanics, Davis and Selvadurai. [337]

- Rheology of the Earth, Ranalli. [1108]

- Deformation of Earth materials, Karato. [715]

- Fundamentals of the Theory of Plasticity, Kachanov. [699]

- Computational methods for plasticity, de Souza Neto et al. [1263]

- Computer simulation of dynamic phenomena, M. wilkins [1457]

- Continuum theory of plasticity, Khan and Huang [738]

- Theory of plasticity, Chakrabarty [234]

- Zienkiewicz Taylor [1541]

- Rheology Principles, Macosko [867]

- Computational Inelasticity, Simo and Hughes [655]

- Lectures on Visco-Plastic Fluid Mechanics, G. Ovarlez & S. Hormozi [1025]

- Complex fluids, P. Saramito [1188]

5.2 Linear viscous aka Newtonian

Simply put, a Newtonian fluid is a fluid in which the viscous stresses at every point are linearly proportional to the local strain rate. Mathematically speaking, this means that the fourth-order tensor $\boldsymbol{C}$ relating the viscous stress tensor to the strain rate tensor does not depend on the stress state and velocity of the flow.

$$\boldsymbol{\tau} = \boldsymbol{C} : \dot{\boldsymbol{\varepsilon}} \tag{5.1}$$

One very often makes the assumption that the fluid is isotropic, i.e. its mechanical properties are the same along any direction. As a consequence the fourth order viscosity tensor $\boldsymbol{C}$ is symmetric and will have only two independent real parameters: a bulk viscosity coefficient, that defines the resistance of the medium to gradual uniform compression; and a dynamic viscosity coefficient η that expresses its resistance to gradual shearing[1].

Rather logically we denote by non-Newtonian fluids which are not Newtonian, i.e. their viscosity (tensor) depends on stress. Such fluids are part of our daily life, e.g. honey, toothpaste, paint, blood, or shampoo. They are also sometimes denoted as Generalized Newtonian Fluid .

no idea where this comes from ...
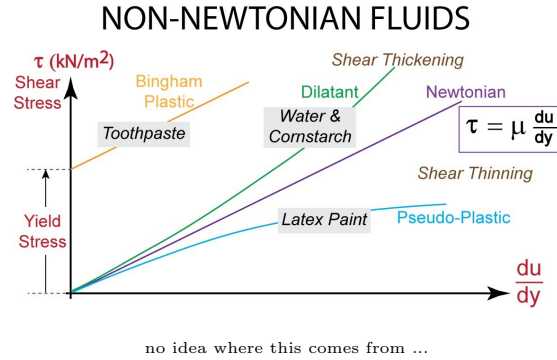
5.3 Power-law model

One of the simplest non-Newtonian viscosity model is the power law model, for which the viscosity depends on the (effective) deviatoric strain rate as follows:

$$\eta(\dot{\varepsilon}_e) = K\dot{\varepsilon}_e^{n-1} \qquad \text{or} \qquad \sigma = 2K\dot{\varepsilon}_e^n \tag{5.2}$$

where n and K are parameters. n is called the power law index. $\dot{\varepsilon}_e$ is defined in (4.73) and in the table here above. Note that a Newtonian viscosity is recovered when $n = 1$. Also n and K may depend on temperature (see Reddy [1118, p339]).

A so-called 'generalised' power law rheology is proposed in Iaffaldano & bunge (2009) [650]:

$$\eta = K(\dot{\varepsilon}_e + \dot{\varepsilon}_0)^{n-1} \tag{5.3}$$

so that in the rigid areas where $\dot{\varepsilon}_e \to 0$ the rheology uses instead a minimum strain rate value $\dot{\varepsilon}_0$.

Relevant Literature: England & Molnar (1997) [400]

5.4 Carreau model

carreau_model.tex

Note that this model is sometimes called Bird-Carreau in the literature. As explained in Reddy [1118], the power-law model poses no restriction on how small or large the viscosity may become, which may prove problematic once implemented as it can lead to runaway effects (strain rate becomes large → viscosity becomes smaller → strain rate becomes larger, etc ...). This problem is alleviated in the so-called Carreau [2] model [226] (see for example Zinani & Frey (2007) [1546]). The viscosity is then given by[3]

$$\eta(\dot{\varepsilon}_e) = \eta_\infty + (\eta_0 - \eta_\infty)\left(1 + (\lambda\dot{\varepsilon}_e)^2\right)^{(n-1)/2} \tag{5.4}$$

---

[1]We here neglect the so-called rotational viscosity coefficient which results from a coupling between the fluid flow and the rotation of the individual particles

[2]https://en.wikipedia.org/wiki/Carreau_fluid

[3]In [1188] we find a slightly different expression where the square exponent is omitted.

where η_0, η_∞, λ and $n \in [0, 1]$ are material parameters. λ is called the relaxation time: it is the inverse of the shear rate at which the fluid changes from Newtonian to power-law behavior. This expression is to be found in the Comsol manual[4].

At low strain rate a Carreau fluid behaves as a Newtonian fluid with viscosity η_0. At intermediate strain rates $\dot{\varepsilon}_e \lambda \sim 1$ a Carreau fluid behaves as a Power-law fluid. At high strain rate, a Carreau fluid behaves as a Newtonian fluid again with viscosity η_∞.
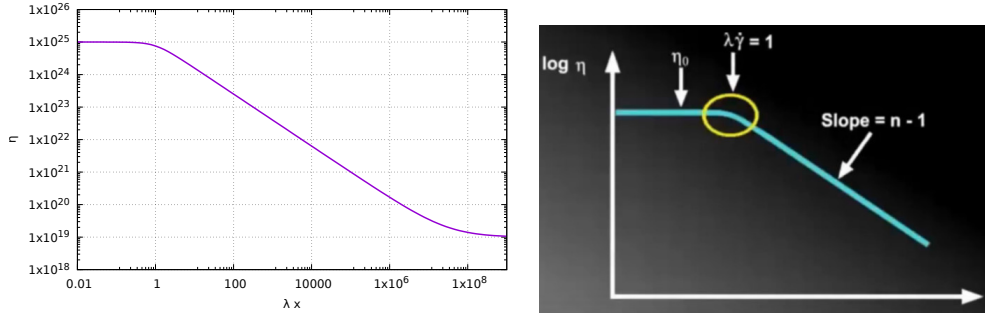
Left: Carreau model effective viscosity as a function of the product $\lambda \dot{\varepsilon}_e$. Right: taken from video at `https://youtu.be/qErs5zZV4BQ`.

Note that the (Bird)-Carreau-Yasuda model [1483, 1022] is very similar to the standard (Bird)-Carreau:

$$\eta = \eta_\infty + (\eta_0 - \eta_\infty) \left(1 + (\lambda\dot{\varepsilon}_e)^a\right)^{(n-1)/a} \tag{5.5}$$

It is for instance used in van de Vosse *et al.* (2003) [1393] to model blood.

Flows in a Lid-Driven Cavity with this rheology are presented in [1546, 1231].

Relevant Literature: Bercovici (1993) [81], Bercovici (1995) [82], Marcotte (2000) [882], Huerta & Liu (1988) [636].

5.5 Bingham model

Bingham [95] fluids can sustain an applied stress without any motion occuring. Only when the applied stress exceeds a yield stress τ_0 then the fluid flows. This translates as follows [1118]:

$$\boldsymbol{\tau} = \left(\frac{\tau_0}{\dot{\varepsilon}} + 2\eta_0\right)\dot{\boldsymbol{\varepsilon}}^d \qquad \text{if } \tau_e > \tau_0 \tag{5.6}$$

$$\boldsymbol{\tau} = \mathbf{0} \qquad \text{if } \tau_e \leq \tau_0 \tag{5.7}$$

When flow occurs, the effective viscosity is then given by:

$$\eta(\dot{\varepsilon}_e) = \frac{\tau_0}{\dot{\varepsilon}_e} + 2\eta_0 \tag{5.8}$$

and when the strain rate is large we recover a Newtonian behaviour. Typical Bingham fluids are mud, slurry, toothpaste.

When using a velocity-based FEM code, the implementation of this rheological behaviour is complicated by the no-flow condition under a given stress. However, our codes require a relationship between stress and strain rate in the form of an effective viscosity which cannot be zero. This difficulty can be circumvented by implementing Bingham fluids as follows [1118]:

$$\boldsymbol{\tau} = \left(\frac{\tau_0(1 - \eta/\eta_r)}{\dot{\varepsilon}_e} + 2\eta_0\right)\dot{\boldsymbol{\varepsilon}} \qquad \text{if } \tau_e > \tau_0 \tag{5.9}$$

$$\boldsymbol{\tau} = 2\eta_r\dot{\boldsymbol{\varepsilon}} \qquad \text{if } \tau_e \leq \tau_0 \tag{5.10}$$

---

[4] `https://doc.comsol.com/5.3/doc/com.comsol.help.cfd/cfd_ug_fluidflow_single.06.062.html`

where η_r is a pre-yield viscosity and $\eta/\eta_r << 1$ (typically 1% or less). This is a form of regularisation, and we will see a similar one in the next section.

Note the interesting paper by Barnes and Walter (1985) [50] who argue that "the yield stress concept is an idealization, and that, given accurate measurements, no yield stress exists. The simple Cross model is shown to be a useful empiricism for many non-Newtonian fluids, including those which have hitherto been thought to possess a yield stress." The Cross model is presented in Section 5.9.

🔖 Relevant Literature: Papanastasiou (1987) [1031], Blackery & Mitsoulis (1997) [101], Mitsoulis & Zisis (2001) [935], Mahmood *et al.* (2017) [870], Syrakos *et al.* (2014) [1307], Bingham [95], Balmforth & Rust (2009) [45], Grinevich & Olshanskii (2009) [526], Sverdrup *et al.* (2018) [1305] FE method for incompressible non-Newtonian flow (Bercovier & Engelman (1980) [85]); Flow around a rigid sphere (Liu *et al.* (2002) [840]), Conduit flow of an incompressible, yield-stress fluid, Taylor and Wilson [1325] (1997).

5.6   Herschel-Bulkley visco-plastic model

The Herschel-Bulkley model is effectively a combination of the power-law model and a simple plastic model:

$$\boldsymbol{\tau} = 2\left(K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}}\right)\dot{\boldsymbol{\varepsilon}} \qquad \text{if } \tau_e > \tau_0 \tag{5.11}$$

$$\dot{\boldsymbol{\varepsilon}} = \mathbf{0} \qquad \text{if } \tau_e \leq \tau_0 \tag{5.12}$$

in which τ_0 is the yield stress, K the consistency, and n is the flow index [90]. The flow index measures the degree to which the fluid is shear-thinning ($n < 1$) or shear-thickening ($n > 1$). If $n = 1$ and $\tau_0 = 0$ the model reduces to the Newtonian model.

The term between parenthesis above is the nonlinear effective viscosity. Concretely, the implementation goes as follows[5]:

$$\eta(\dot{\boldsymbol{\varepsilon}}) = \begin{cases} \eta_0 & \dot{\varepsilon}_e \leq \dot{\varepsilon}_0 \\ K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_e} & \dot{\varepsilon}_e \geq \dot{\varepsilon}_0 \end{cases} \tag{5.13}$$

The limiting viscosity η_0 is chosen such that $\eta_0 = K\dot{\varepsilon}_0^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_0}$

A large limiting viscosity means that the fluid will only flow in response to a large applied force. This feature captures the Bingham-type behaviour of the fluid. Note that when strain rates are large, the power-law behavior dominates.

As we have seen for Bingham fluids, the equations above are not easily amenable to implementation so that one usually resorts to regularisation, which is a modification of the equations by introducing a new material parameter which controls the exponential growth of stress. This way the equation is valid for both yielded and unyielded areas (Blackery & Mitsoulis (1997) [101], Papanastasiou (1987) [1031], Zinani & Frey (2007) [1546], Sverdrup *et al.* (2018) [1305]):

$$\eta(\dot{\varepsilon}_e) = K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_e}[1 - \exp(-m\dot{\varepsilon}_e)] \tag{5.14}$$

When the strain rate becomes (very) small a Taylor expansion of the regularisation term yields $1 - \exp(-m\dot{\varepsilon}) \sim m\dot{\varepsilon}$ so that $\eta_{eff} \to m\tau_0$. However, it seems more physically meaningful to replace m by a reference strain rate value $\dot{\varepsilon}_0$ so that

$$\eta_{eff}(\dot{\boldsymbol{\varepsilon}}) = K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_e}\left[1 - \exp\left(-\frac{\dot{\varepsilon}_e}{\dot{\varepsilon}_0}\right)\right] \tag{5.15}$$

---

[5]`https://en.wikipedia.org/wiki/Herschel-Bulkley_fluid`

In this case, when strain rate becomes (very) small a Taylor expansion of the regularisation term yields

$$\frac{\tau_0}{\dot{\varepsilon}_e}\left[1 - \exp\left(-\frac{\dot{\varepsilon}_e}{\dot{\varepsilon}_0}\right)\right] \simeq \frac{\tau_0}{\dot{\varepsilon}_e}\frac{\dot{\varepsilon}_e}{\dot{\varepsilon}_0} = \frac{\tau_0}{\dot{\varepsilon}_0} \tag{5.16}$$

This has the dimensions of a viscosity and this is effectively the definition of a maximum viscosity η_{max}.

🔖 Relevant Literature:

- Viscous flow with large free surface motion (Huerta & Liu (1988) [636]);

- Numerical simulation of thermal plumes (Massmeyer *et al.* [889]);

- Flows Through a Sudden Axisymmetric Expansion (Machado *et al.* [864], Jay *et al.* (2001) [678]);

- Dam break problem (Ancey & Cochard (2009) [17], Cochard & Ancey (2009) [280], Balmforth *et al.* [44];

- Weakly compressible Poiseuille flow (Taliadorou (2009) [1316]);

- Flow past cylinders in tubes (Mitsoulis & Galazoulas (2009) [934]);

- Determination of yield surfaces (Burgos & Alexandrou (1999) [194]);

- Carbopol hydrogel rheology for experimental tectonics and geodynamics Di Giuseppe et al. [353] (2015);

- Flow past a sphere(disc) Besses, Magnin, and Jay [90] (2004); Gavrilov *et al.* (2017) [470]).

- Progress in numerical simulation of yield stress fluid flows, Saramito and Wachs [1189] (2017);

- elastoviscoplastic model based on the Herschel–Bulkley viscoplastic model Saramito [1187] (2019).

5.7 The Casson model

It is described in Barnes (1999) [49]:

$$\sqrt{\sigma} = \sqrt{\sigma_y} + \sqrt{\eta_p \dot{\varepsilon}_e} \tag{5.17}$$

or, when squaring it:

$$\sigma = \sigma_y + \eta_p \dot{\varepsilon}_e + 2\sqrt{\sigma_p \eta_p \dot{\varepsilon}_e} \tag{5.18}$$

This model has been found to accurately describe the behaviour of synthetic based muds [2]. See also Section 2.5.1 of Macosko [867].

5.8 The Ellis model

An Ellis equation would be of the form [1146]

$$\frac{\eta - \eta_\infty}{\eta_0 - \eta_\infty} = \frac{1}{1 + (\sigma/\sigma_c)^m} \tag{5.19}$$

where σ is the shear stress, σ_c is a critical shear stress and m is a large number. See also Section 2.4.3 of Macosko [867].

5.9 One model to rule them all?

Let us consider the base equation

$$\boxed{\frac{\eta - \eta_\infty}{\eta_0 - \eta_\infty} = [1 + (K\dot{\varepsilon}_e)^a]^{-(1-n)/a}} \tag{5.20}$$

This equation is purposefully generic and specific parameter combination choices allow to recover any of the above models (and more) [1022]. See also an early paper by Cross (1965) [307] for a somewhat similar equation. See also Section 2.4.2 of Macosko [867].

Similar conclusions are reached in the following video:

https://youtu.be/dVCb11dZR7Y

Taken from Huerta and Liu [636] (1988)

5.10 Dislocation and Diffusion creep

insert here background and links to relevant textbooks

The standard dislocation creep effective viscosity is given by:

$$\eta^{ds}(p, T, \dot{\varepsilon}) = \eta^{ds}(p, T, \dot{\varepsilon}_e) = \frac{1}{2} f A^{-1/n} \dot{\varepsilon}_e^{(1-n)/n} \exp\left(\frac{Q + pV}{nRT}\right)$$

where A is the pre-exponential scaling factor, f is a scaling factor representing viscous weakening or strengthening, Q is the activation energy, V is the activation volume, T is the absolute temperature, n is the power-law exponent, R is the universal gas constant.

The coefficients A, n, Q, V are material parameters and are obtained in the laboratory by means of high pressure/temperature experiments (see for instance Karato & Wu (1993) [714]). Unfortunately these experiments cannot be run at Earth-like strain rate values ($\sim 10^{-15}\text{s}^{-1}$) so that extrapolations must be carried out over several orders of magnitude to arrive at values we can use in our numerical models. The 1/2 factor arises from the relationship between deviatoric stress and strain rate which involves a factor 2.

The factor f is in fact a tuning parameter used to explore end members (e.g. 'weak crust' vs 'strong crust'), see discussion in the supplementary material in Huismans & Beaumont (2011) [645].

This approach has been extensively used by the SOPALE users community, see for instance Warren *et al.* (2008) [1442, 1443, 1444] or Gray & Pysklywec (2012) [511].

<div style="background: orange;">insert here equation for diffusion creep</div>

Furthermore, we know that several other factors will strongly affect the rheology:

- water content, or as often mentioned: 'dry' vs 'wet'. Following [714], dry means water-free and wet means water-saturated conditions.

Mechanism	Dry	Wet
Dislocation creep		
A (s^{-1})	3.5×10^{22}	2.0×10^{18}
n	3.5	3.0
m	0	0
E^* (kJ mol^{-1})	540	430
V^* (cm^3 mol^{-1})	15 to 25†	10 to 20†
Diffusion creep		
A (s^{-1})	8.7×10^{15}	5.3×10^{15}
n	1.0	1.0
m	2.5‡	2.5‡
E^* (kJ mol^{-1})	300	240
V^* (cm^3 mol^{-1})	6§	5§

†The activation volume for dislocation creep is not well constrained. Values from 13 cm³ mol⁻¹ at wet conditions to 27 cm³ mol⁻¹ at nearly dry conditions have been reported (*33*). Considering this uncertainty, we used a range of activation volumes, 15 to 25 cm³ mol⁻¹ for dry olivine and 10 to 20 cm³ mol⁻¹ for wet olivine.   ‡The grain-size exponent is reported as 2 for dry olivine and 3 for wet olivine (*16*). However, during extrapolation to a large grain size, the grain-size exponent may change from 2 to 3 or from 3 to 2 (*16*). Taking this uncertainty into account, we chose a grain-size exponent of 2.5, and the preexponential factors were modified accordingly.   §The activation volume for diffusion is determined as 6 cm³ mol⁻¹ at dry conditions (*26*). No data are available for the activation volume for diffusion at wet conditions; we assumed that this value is ~80% of the activation volume at dry conditions.

Taken from Karato and Wu [714].

📚 Relevant Literature: Quinquis & Buiter (2014) [1093] and refs therein for the effects of water migration on models of subduction dynamics.

- composition: while one typically assigns olivine properties to the mantle in models, the mineral olivine[6] is actually a magnesium iron silicate with the formula $(Mg^{2+}, Fe^{2+})_2SiO_4$. and the ratio of magnesium to iron varies between the two endmembers of the solid solution series: forsterite (Mg-endmember: Mg_2SiO_4) and fayalite (Fe-endmember: Fe_2SiO_4).

- grain size: this only affects diffusion creep mechanisms [714]. Grain size varies over several orders of magnitude and also evolves over time and its evolution is affected by the ambient deformation and the deformation history. Dannberg *et al.* [320] then used a diffusion creep effective viscosity given by:

$$\eta^{df} = \frac{1}{2} A_{df}^{-1} d^m \exp\left(\frac{Q_{df} + pV_{df}}{RT}\right)$$

where d is the (variable) grain size and m the grain size exponent. Grain growth/evolution is usually approximated using semi-empirical expressions [320, section 2.2]. Smaller grains facilitating faster creep.

Relevant literature on this topic is in Section **??**.

- anisotropy, LPO: see relevant literature in Section **??**.

- phase changes

Remark. *It is not uncommon to find in the literature effective viscosity formulations written as a function of B with $B = A^{-1/n}$ [1442, 1443, 1444]. Also, this B coefficient often contains the conversion factor of the next remark.*

Remark. *Material parameters obtained in the lab are often measured on a uniaxial machine. An additional coefficient is added to the effective viscosity formula (see [511, 512], or Table 1a of Warren et al. (2008) [1442]): $3^{-(1+n)/2n}2^{(1-n)/n}$. See page 77 of Ranalli [1108] for an explanation.*

---

[6]https://en.wikipedia.org/wiki/Olivine

Remark. *In Tullis, Horowitz, and Tullis [1377] it is explained how to formulate a flow law for a polyphase aggregate from end-member flow laws.*
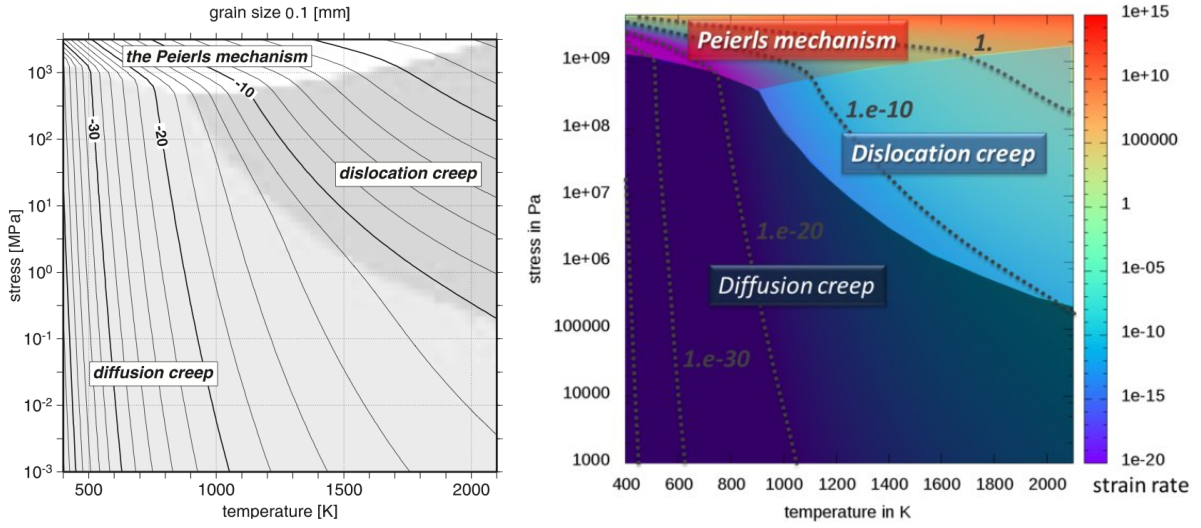
Fig-B: Left: Taken from Kameyama *et al.* (1999) [704]. Deformation mechanism map calculated for grain size $a = 0.1$mm. The lightly shaded area indicates that deformation mainly occurs by diffusion creep. The densely shaded area indicates that deformation mainly occurs by power-law creep. The white region indicates that deformation mainly occurs by the Peierls mechanism. The solid curves are lines of constant strain rate. The numbers attached to each contour indicate the logarithm of the strain rate in the unit of s^{-1}. Right: Taken from Elbeshausen & Melosh (1998) [389]. Strain rate as a function of stress and temperature. The parameter space dominated by each of the terms is denoted by different hues, depending on strain rate. Note that the original equations implicitly include a pressure dependence in the enthalpy term, which is not strong and therefore neglected here. The diffusion regime is highly temperature dependent and important for small strain rates only. Dislocation creep occurs for intermediate stresses and higher temperatures, while the Peierls mechanism dominates at higher stresses (≥ 500MPa) and shows a strong stress dependence for low temperatures. The dashed contours are strain rate in units of s^{-1}.

A closer look at the diffusion creep of Karato & Wu (1993)   In the article, the following equation is used:

$$\dot{\varepsilon} = A \left(\frac{\tau}{\mu} \right) \left(\frac{b}{d} \right)^m \exp \left(- \frac{Q + pV}{RT} \right)$$

where μ is the shear modulus ($\sim$80GPa), b is the length of the Burgers vector ($\sim$0.5nm) and d is the grain size. One can express the above equation in terms of second invariants (see Section 4.3):

$$\dot{\underline{\varepsilon}}_e = A \left(\frac{\tau_e}{\mu} \right) \left(\frac{b}{d} \right)^m \exp \left(- \frac{Q + pV}{RT} \right)$$

and assuming a Newtonian linearisation/relation between deviatoric stress and strain rate $\underline{\tau}_e = 2\eta^{df} \underline{\dot{\epsilon}}_e$, one arrive at

$$\eta^{df} = \frac{1}{2} \left(\frac{A}{\mu} \right)^{-1} \left(\frac{b}{d} \right)^{-m} \exp \left(\frac{Q + pV}{RT} \right)$$

or,

$$\eta^{df} = \frac{1}{2} \left[\frac{A}{\mu} \left(\frac{b}{d} \right)^m \right]^{-1} \exp \left(\frac{Q + pV}{RT} \right)$$

The effective diffusion creep viscosity is independent of strain-rate so that one could substitute the total pressure for lithostatic pressure in the equation, assume a geotherm and easily compute the predicted viscosity as a function of grain size d.

Let us assume that the 1D profile starts from the base of the lithosphere (say 120km depth) and ends at the 660 boundary. Assume the temperature to increase linearly from 1300 °C to T_{bottom} (to be specified). At the bottom of the lithosphere, the lithostatic pressure is of the order of $\rho \cdot$

$g \cdot L \simeq 3000 \cdot 10 \cdot 120e3 \simeq 4\text{GPa}$. At the bottom of the domain, the pressure has increased by $3300 \cdot 10 \cdot 630e3 \simeq 21\text{GPa}$.

The viscosity profile is plotted hereunder for three different grain sizes, bottom temperature and activation volumes ($4,5,6$ cm^3/mol).
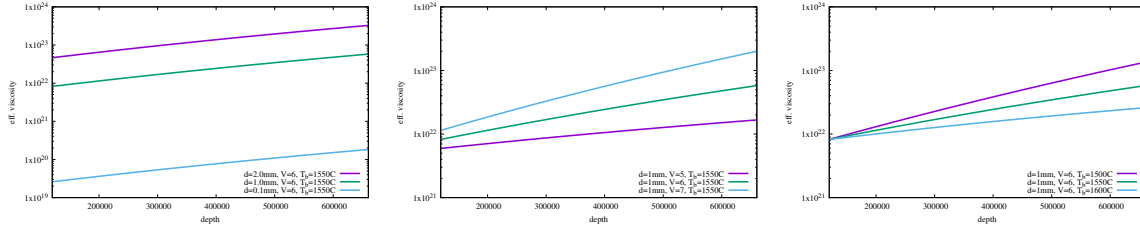
Effective diffusion creep viscosity for various grain size, activation volume and basal temperature values.

images/rheology/kawudiff/

Although this exercise only provides us with first-order results, we can conclude that one can essentially change the diffusion creep effective viscosity by up to 2 orders of magnitude simply by choosing key parameters within acceptable ranges.

📖 Relevant Literature: Dixon and Durham [356] "Measurement of activation volume for creep of dry olivine at upper-mantle conditions"

5.11 The von Mises failure criterion

The von Mises yield criterion suggests that the yielding of materials begins when the second deviatoric stress invariant $\mathcal{I}_2(\boldsymbol{\tau})$ reaches a critical value. For this reason, it is sometimes called the J_2-plasticity or J_2 flow theory[7]. It is part of a plasticity theory that applies best to ductile materials, such as metals.

In material science and engineering the von Mises yield criterion can be also formulated in terms of the von Mises stress or equivalent tensile stress, σ_v, a scalar stress value that can be computed from the stress tensor. In this case, a material is said to start yielding when its von Mises stress reaches a critical value known as the yield strength, σ_Y. The von Mises stress is used to predict yielding of materials under any loading condition from results of simple uniaxial tensile tests. The von Mises stress satisfies the property that two stress states with equal distortion energy have equal von Mises stress.

Because the von Mises yield criterion is independent of the first stress invariant, $\mathcal{I}_1(\boldsymbol{\sigma})$, it is applicable for the analysis of plastic deformation for ductile materials such as metals, as the onset of yield for these materials does not depend on the hydrostatic component of the stress tensor.

Although formulated by Maxwell in 1865, it is generally attributed to von Mises [925]. Huber (1904), in a paper in Polish, anticipated to some extent this criterion. Heinrich Hencky formulated the same criterion as von Mises independently in 1924 [595, 1321]. This criterion is also referred to as the Maxwell-Huber-Hencky-von Mises theory.

The von Mises yield criterion (also known as Prandtl-Reuss yield criterion) is expressed in the principal stresses as

$$\sqrt{3}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} = c$$

or, using the definition of $\mathcal{I}_2(\boldsymbol{\tau})$ as a function of $\sigma_{1,2,3}$ (see Eq. (4.75)):

$$\frac{1}{2}[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2] = c^2$$

---

[7]J_2 is the common notation for $\mathcal{I}_2(\boldsymbol{\tau})$.

where c is the yield stress in uniaxial tension. Note that the $\sqrt{3}$ is necessary so as to match the Tresca yield envelope for an identical c value. It is present also in `https://en.wikipedia.org/wiki/Von_Mises_yield_criterion` CHECK/NEW

The von Mises yield criterion writes:

$$F^{\text{VM}} = \sqrt{3}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - c \tag{5.21}$$

which is very similar to the Drucker-Prager criterion with $\phi = 0$ (see Section 5.14).

The following figure shows the von Mises yield surface in the three-dimensional space of principal stresses (see STONE 159).
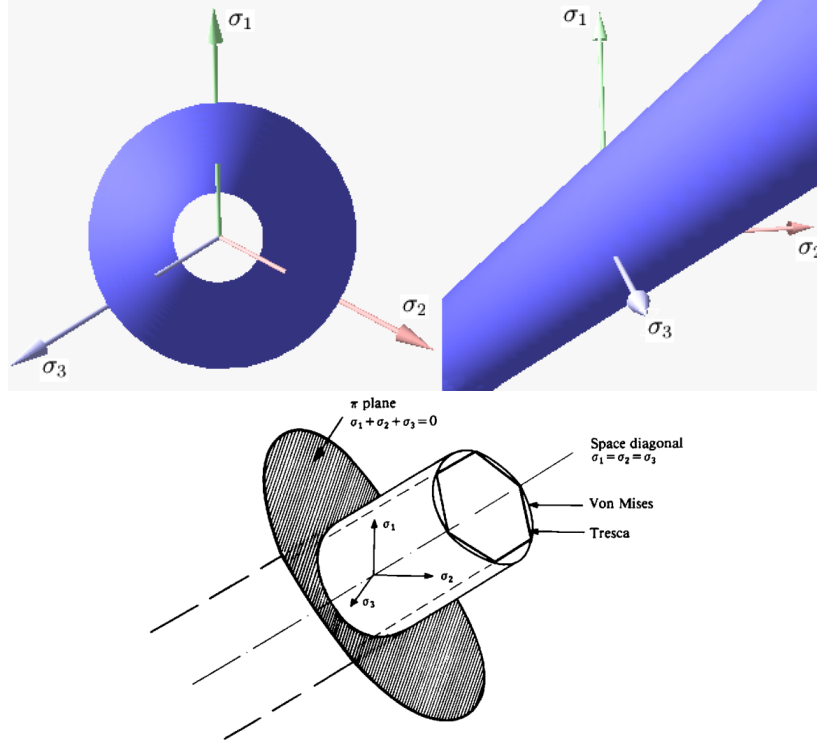
Left: Taken from `https://en.wikipedia.org/wiki/Yield_surface`. Right: Taken from Owen and Hinton [1026].

It is a circular cylinder of infinite length with its axis inclined at equal angles to the three principal stresses.

📥 Relevant Literature: Tasos C Papanastasiou. "Flows of materials with yield". In: *Journal of Rheology* 31.5 (1987), pp. 385–404, F. Tin-Loi and N.S. Ngo. "Performance of the p-version finite element method for limit analysis". In: *International Journal of Mechanical Sciences* 45 (2003), pp. 1149–1166. DOI: `10.1016/j.ijmecsci.2003.08.004`

In the coming paragraph sqrt3 is not taken into account yet

The yield surface Let us try to draw the yield function in the space $\sigma_1, \sigma_2, \sigma_3$. It is given by

$$\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} = c \tag{5.22}$$

$$\Rightarrow \quad \mathcal{I}_2(\boldsymbol{\tau}) = c^2 \tag{5.23}$$

$$\Rightarrow \quad \frac{1}{6}\left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2\right] = c^2 \tag{5.24}$$

$$\Rightarrow \quad (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2 = 6c^2 \tag{5.25}$$

or, temporarily setting $x = \sigma_1$, $y = \sigma_2$ and $z = \sigma_3$:

$$(x-y)^2 + (y-z)^2 + (x-z)^2 = 6c^2 \tag{5.26}$$

$$(x-y)^2 + y^2 - 2yz + z^2 + x^2 - 2xz + z^2 = 6c^2 \tag{5.27}$$

$$2z^2 - 2(x+y)z + (x-y)^2 + x^2 + y^2 - 6c^2 = 0 \tag{5.28}$$

If we wish to plot the envelope in x, y, z space, we need to find (for example) a relation $z(x, y)$. We see immediately that this equation is a second order polynomial in z. Its discriminant Δ is

$$
\begin{aligned}
\Delta &= 4(x+y)^2 - 4 \cdot 2 \cdot [(x-y)^2 + x^2 + y^2 - 6c^2] \\
&= 4x^2 + 8xy + 4y^2 - 8[x^2 - 2xy + y^2 + x^2 + y^2 - 6c^2] \\
&= 4x^2 + 8xy + 4y^2 - 8[2x^2 - 2xy + 2y^2 - 6c^2] \\
&= 4x^2 + 8xy + 4y^2 - 16x^2 + 16xy - 16y^2 + 48c^2 \\
&= -12x^2 + 24xy - 12y^2 + 48c^2 \\
&= -12(x^2 - 2xy + y^2) + 48c^2 \\
&= -12(x-y)^2 + 48c^2
\end{aligned}
$$

Since I am looking for $z(x, y) \in \mathbb{R}$ then $\Delta > 0$ and this imposes a restriction on admissible x, y pairs:

$$
-12(x-y)^2 + 48c^2 > 0
$$

$$
(x-y)^2 < 4c^2
$$

$$
x - y < 2c \qquad \text{or,} \qquad y - x < 2c
$$

$$
y > x - 2c \qquad \text{or,} \qquad y < x + 2c
$$

So the discriminant is positive in the band given by $y > x - 2c$ and $y < x + 2c$ in the x, y-plane, which is a band centered around the line $y = x$. When $\Delta > 0$ we have then

$$
z = \frac{2(x+y) \pm \sqrt{\Delta}}{4}
$$

which means that for each pair x, y there are 2 z values. The middle of this surface is given by the line $z = (x+y)/2$. The plane normal to this line is given by $z = -2(x+y)$.

This approach is reasonably simple for the von Mises criterion but quickly becomes intractable for other criteria. implement in f159? try GM ? DP ?

We now look into the derivatives of the von Mises plastic potential $Q^{\text{vM}}(\boldsymbol{\sigma})$. We have

$$
Q^{\text{vM}}(\boldsymbol{\sigma}) = \sqrt{3}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - c \tag{5.29}
$$

Then

$$
\frac{\partial Q^{\text{vM}}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} = 0 \tag{5.30}
$$

$$
\frac{\partial Q^{\text{vM}}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} = \sqrt{3} \tag{5.31}
$$

$$
\frac{\partial Q^{\text{vM}}}{\partial \theta_{\text{L}}(\boldsymbol{\tau})} = 0 \tag{5.32}
$$

so

$$
C_1^{\text{vM}} = 0 \tag{5.33}
$$

$$
C_2^{\text{vM}} = \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}(1 - 0) = \sqrt{3}\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \tag{5.34}
$$

$$
C_3^{\text{vM}} = 0 \tag{5.35}
$$

124

5.12 The Tresca failure criterion

The Tresca or maximum shear stress yield criterion is taken to be the work of Henri Tresca. It is also referred as the Tresca-Guest (TG) criterion. The functional form of this yield criterion is

$$f(\sigma_1, \sigma_2, \sigma_3) = 0$$

In terms of the principal stresses the Tresca criterion is expressed as

$$\max(|\sigma_1 - \sigma_2|, |\sigma_2 - \sigma_3|, |\sigma_3 - \sigma_1|) = c$$

The following figure shows the Tresca-Guest yield surface in the three-dimensional space of principal stresses.
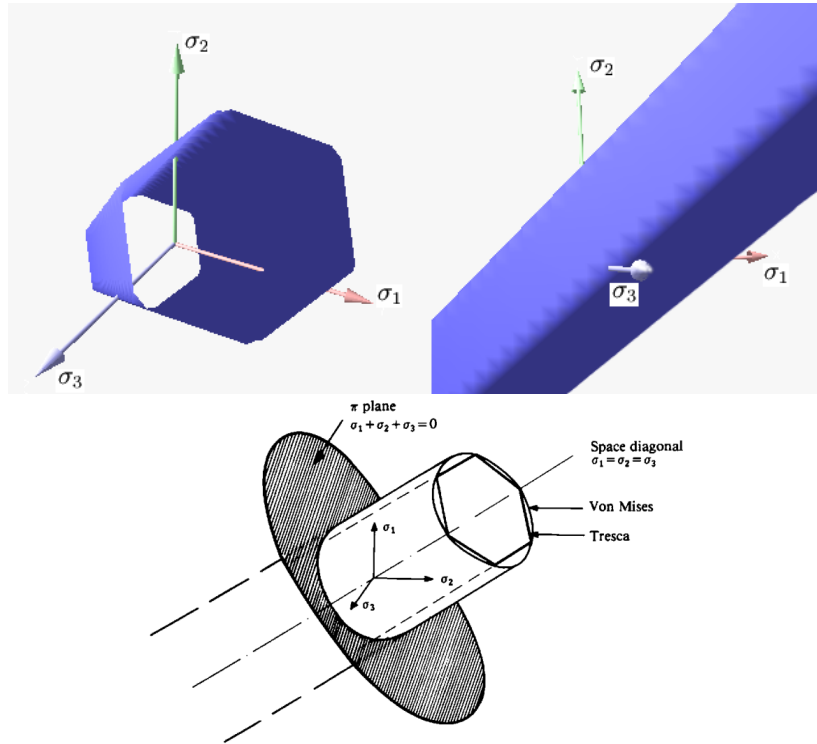
Left: Taken from https://en.wikipedia.org/wiki/Yield_surface. Right: Taken from Owen and Hinton [1026].

It is a prism of six sides and having infinite length. This means that the material remains viscous when all three principal stresses are roughly equivalent (a hydrostatic pressure), no matter how much it is compressed or stretched. However, when one of principal stresses becomes smaller (or larger) than the others the material is subject to shearing. In such situations, if the shear stress reaches the yield limit then the material enters the plastic domain.

Remark. *The yield function presents sharp corners, making its numerical implementation more difficult (directional derivatives are needed)*

We have already established in Eq. (4.58):

$$\sigma_1 - \sigma_3 \;=\; 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\cos\theta$$

with $\sigma_1 > \sigma_2 > \sigma_3$, so that the failure criterion is given by

$$F^{\mathrm{TR}} = 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\cos\theta - c$$

We now look into the derivatives of the plastic potential $Q^{\mathrm{Tr}}(\boldsymbol{\sigma})$. We have

$$Q^{\mathrm{Tr}} = 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\cos\theta_{\mathrm{L}}(\boldsymbol{\tau}) - c$$

$$\frac{\partial Q^{\mathrm{Tr}}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} = 0 \tag{5.36}$$

$$\begin{aligned}
\frac{\partial Q^{\mathrm{Tr}}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} &= 2\cos\theta_{\mathrm{L}} - 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}}\frac{\partial\theta_{\mathrm{L}}}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \\
&= 2\cos\theta_{\mathrm{L}} - 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}}\frac{\partial\theta_{\mathrm{L}}}{\partial\mathcal{I}_2(\boldsymbol{\tau})}\frac{\partial\mathcal{I}_2(\boldsymbol{\tau})}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \\
&= 2\cos\theta_{\mathrm{L}} - 2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}}\left(-\frac{1}{2}\tan 3\theta_{\mathrm{L}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})}\right)2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \\
&= 2\cos\theta_{\mathrm{L}} + 2\sin\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}} \\
&= 2\cos\theta_{\mathrm{L}}(1 + \tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) \tag{5.37}
\end{aligned}$$

$$\frac{\partial Q^{\mathrm{Tr}}}{\partial\theta_{\mathrm{L}}(\boldsymbol{\tau})} = -2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}} \tag{5.38}$$

so

$$C_1^{\mathrm{Tr}} = 0 \tag{5.39}$$

$$\begin{aligned}
C_2^{\mathrm{Tr}} &= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(\frac{\partial Q}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} - \frac{\tan 3\theta_{\mathrm{L}}}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\frac{\partial Q}{\partial\theta_{\mathrm{L}}(\boldsymbol{\tau})}\right) \\
&= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(2\cos\theta_{\mathrm{L}}(1 + \tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{\tan 3\theta_{\mathrm{L}}}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}}\right) \\
&= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(2\cos\theta_{\mathrm{L}}(1 + \tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + 2\tan 3\theta_{\mathrm{L}}\sin\theta_{\mathrm{L}}\right) \\
&= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(2\cos\theta_{\mathrm{L}}(1 + \tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + 2\cos\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}\tan\theta_{\mathrm{L}}\right) \\
&= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(2\cos\theta_{\mathrm{L}}(1 + 2\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}})\right) \tag{5.40}
\end{aligned}$$

$$\begin{aligned}
C_3^{\mathrm{Tr}} &= -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}\frac{\partial Q}{\partial\theta_{\mathrm{L}}(\boldsymbol{\tau})} \\
&= -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}(-2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\sin\theta_{\mathrm{L}}) \\
&= \frac{\sqrt{3}}{\mathcal{I}_2(\boldsymbol{\tau})}\frac{\sin\theta_{\mathrm{L}}}{\cos 3\theta_{\mathrm{L}}} \tag{5.41}
\end{aligned}$$

5.13   The Mohr-Coulomb failure criterion

mccriterion.tex

Mohr-Coulomb theory is a model describing the response of a material such as rubble piles or concrete to shear stress as well as normal stress. Most of the classical engineering materials somehow follow this rule in at least a portion of their shear failure envelope. In geology it is used to define shear strength of soils at different effective stresses [560].

In structural engineering it is used to determine failure load as well as the angle of fracture of a displacement fracture in concrete and similar materials. Coulomb's friction hypothesis is used to determine the combination of shear and normal stress that will cause a fracture of the material. Mohr's circle is used to determine the principal stresses that will produce this combination of shear and normal stress, and the angle of the plane in which this will occur. According to the principle of normality, the stress introduced at failure will be perpendicular to the line describing the fracture condition.

(a) Geometrical representation of the Mohr–Coulomb and Drucker–
Prager yield surfaces in principal stress space.
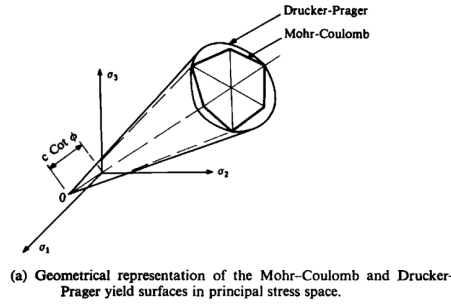
Taken from Owen and Hinton [1026].

The Mohr-Coulomb failure criterion represents the linear envelope that is obtained from a plot of the shear strength of a material versus the applied normal stress. This relation is expressed as (Owen & Hinton book [1026, p219])

$$\tau_m = -\sigma_m \sin\phi + c\cos\phi \tag{5.42}$$

where τ_m is the magnitude of the shear stress, σ_m is the normal stress, c is the intercept of the failure envelope with the τ axis, and ϕ is the slope of the failure envelope. The minus sign in the above equation is for the case where compression is assumed to be negative[8]. The quantity c is often called the cohesion and the angle ϕ is called the angle of internal friction.

We have

$$\tau_m = \frac{\sigma_1 - \sigma_3}{2} \qquad\qquad \sigma_m = \frac{\sigma_1 + \sigma_3}{2}$$

with σ_1 is the maximum principal stress and σ_3 is the minimum principal stress, or

$$\frac{\sigma_1 - \sigma_3}{2} = -\frac{\sigma_1 + \sigma_3}{2}\ \sin\phi + c\ \cos\phi \tag{5.43}$$

Using Eqs. (4.58) and (4.59) for $(\sigma_1 - \sigma_3)/2$ and $(\sigma_1 + \sigma_3)/2$ we get[9]:

$$\frac{\sigma_1 - \sigma_3}{2} = -\frac{\sigma_1 + \sigma_3}{2}\sin\phi + c\cos\phi$$

$$\Rightarrow \quad \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\cos\theta = -\left(\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) - \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\frac{1}{\sqrt{3}}\sin\theta\right)\sin\phi + c\cos\phi$$

$$\Rightarrow \quad \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta - \frac{1}{\sqrt{3}}\sin\theta\sin\phi\right) - c\cos\phi = 0$$

---

[8] https://en.wikipedia.org/wiki/Mohr-Coulomb_theory
[9] This is Eq. (7.16) of Owen and Hinton [1026]

$$F^{\text{MC}} = \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta - \frac{1}{\sqrt{3}}\sin\theta\sin\phi\right) - c\cos\phi \tag{5.44}$$

This formula (without the cohesion) is used in Willett [1459]. Since $p = -\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})$, we also have:

$$F^{\text{MC}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta - \frac{1}{\sqrt{3}}\sin\theta\sin\phi\right) - (p\sin\phi + c\cos\phi) \tag{5.45}$$

Remark. *The expression for F in the Mohr-Coulomb case in Zienkiewicz & Cormeau (1974) [1533] contains errors which are later corrected in Zienkiewicz, Taylor, and Fox [1543, p102].*

📘 Relevant Literature: this criterion is also used in computer graphics animation [1530]

<div style="float:right; background:orange; padding:4px;">when $\phi = 0$ we should recover Tresca but factor 2 is wrong ?</div>

We now look into the derivatives of the Drucker-Prager plastic potential $Q^{\text{MC}}(\boldsymbol{\sigma})$. We have

$$Q^{\text{MC}} = \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta_{\text{L}}(\boldsymbol{\tau}) - \frac{1}{\sqrt{3}}\sin\theta_{\text{L}}(\boldsymbol{\tau})\sin\phi\right) - c\cos\phi$$

Then

$$\frac{\partial Q^{\text{MC}}}{\partial\mathcal{I}_1(\boldsymbol{\sigma})} = \frac{1}{3}\sin\phi \tag{5.46}$$

$$\frac{\partial Q^{\text{MC}}}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} = \cos\theta_{\text{L}} - \frac{1}{\sqrt{3}}\sin\theta_{\text{L}}\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\sin\theta_{\text{L}} - \frac{1}{\sqrt{3}}\cos\theta_{\text{L}}\sin\phi\right)\frac{\partial\theta_{\text{L}}}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}$$

$$= \cos\theta_{\text{L}} - \frac{1}{\sqrt{3}}\sin\theta_{\text{L}}\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\sin\theta_{\text{L}} - \frac{1}{\sqrt{3}}\cos\theta_{\text{L}}\sin\phi\right)\frac{\partial\theta_{\text{L}}}{\partial\mathcal{I}_2(\boldsymbol{\tau})}\frac{\partial\mathcal{I}_2(\boldsymbol{\tau})}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}$$

$$= \cos\theta_{\text{L}} - \frac{1}{\sqrt{3}}\sin\theta_{\text{L}}\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\sin\theta_{\text{L}} - \frac{1}{\sqrt{3}}\cos\theta_{\text{L}}\sin\phi\right)\left(-\frac{1}{2}\tan 3\theta_{\text{L}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})}\right)2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$$

$$= \cos\theta_{\text{L}} - \frac{1}{\sqrt{3}}\sin\theta_{\text{L}}\sin\phi + \left(\sin\theta_{\text{L}} + \frac{1}{\sqrt{3}}\cos\theta_{\text{L}}\sin\phi\right)\tan 3\theta_{\text{L}}$$

$$= \cos\theta_{\text{L}}\left[1 - \frac{1}{\sqrt{3}}\tan\theta_{\text{L}}\sin\phi + \left(\tan\theta_{\text{L}} + \frac{1}{\sqrt{3}}\sin\phi\right)\tan 3\theta_{\text{L}}\right]$$

$$= \cos\theta_{\text{L}}\left[(1 + \tan\theta_{\text{L}}\tan 3\theta_{\text{L}}) + \frac{1}{\sqrt{3}}\sin\phi(\tan 3\theta_{\text{L}} - \tan\theta_{\text{L}})\right] \tag{5.47}$$

$$\frac{\partial Q^{\text{MC}}}{\partial\theta_{\text{L}}(\boldsymbol{\tau})} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(-\sin\theta_{\text{L}} - \frac{1}{\sqrt{3}}\cos\theta_{\text{L}}\sin\phi\right)$$

$$= -\frac{1}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}(\sqrt{3}\sin\theta_{\text{L}} + \cos\theta_{\text{L}}\sin\phi) \tag{5.48}$$

so

$$C_1^{\mathrm{MC}} \;=\; \frac{1}{3}\sin\phi$$

$$C_2^{\mathrm{MC}} \;=\; \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(\frac{\partial Q}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} - \frac{\tan 3\theta_{\mathrm{L}}}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\frac{\partial Q}{\partial\theta_{\mathrm{L}}(\boldsymbol{\tau})}\right)$$

$$= \; \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(\cos\theta_{\mathrm{L}}\left[(1+\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{1}{\sqrt{3}}\sin\phi(\tan 3\theta_{\mathrm{L}} - \tan\theta_{\mathrm{L}})\right] + \frac{\tan 3\theta_{\mathrm{L}}}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\frac{1}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}(\sqrt{3}\sin\right.$$

$$= \; \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(\cos\theta_{\mathrm{L}}\left[(1+\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{1}{\sqrt{3}}\sin\phi(\tan 3\theta_{\mathrm{L}} - \tan\theta_{\mathrm{L}})\right] + \tan 3\theta_{\mathrm{L}}(\sin\theta_{\mathrm{L}} + \frac{1}{\sqrt{3}}\cos\theta_{\mathrm{L}}\,\mathrm{s}\right.$$

$$= \; \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\left(\cos\theta_{\mathrm{L}}\left[(1+\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{1}{\sqrt{3}}\sin\phi(\tan 3\theta_{\mathrm{L}} - \tan\theta_{\mathrm{L}}) + \tan 3\theta_{\mathrm{L}}(\tan\theta_{\mathrm{L}} + \frac{1}{\sqrt{3}}\sin\phi)\right]\right)$$

$$= \; \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\cos\theta_{\mathrm{L}}\left[(1+2\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{1}{\sqrt{3}}\sin\phi(2\tan 3\theta_{\mathrm{L}} - \tan\theta_{\mathrm{L}})\right]$$

$$C_3^{\mathrm{MC}} \;=\; -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}\frac{\partial Q}{\partial\theta_{\mathrm{L}}(\boldsymbol{\tau})}$$

$$= \; -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}}\frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}\left[-\frac{1}{\sqrt{3}}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}(\sqrt{3}\sin\theta_{\mathrm{L}} + \cos\theta_{\mathrm{L}}\sin\phi)\right]$$

$$= \; \frac{\sqrt{3}\sin\theta_{\mathrm{L}} + \sin\phi\cos\theta_{\mathrm{L}}}{2\mathcal{I}_2(\boldsymbol{\tau})\cos 3\theta_{\mathrm{L}}}$$

5.14　The Drucker-Prager failure criterion

　　The von Mises yield criterion is not suitable for modelling the yielding of frictional material as it does not include the effect of mean stress as observed in experiments. To overcome this limitation, Drucker and Prager (1952) [370] proposed a revised function for frictional materials.

　　The Drucker-Prager yield criterion has the function form

$$\mathcal{F}^{\mathrm{DP}}(\boldsymbol{\sigma}) = \mathcal{F}\left(\mathcal{I}_1(\boldsymbol{\sigma}), \mathcal{I}_2(\boldsymbol{\tau})\right) = 0 \tag{5.51}$$

This criterion is most often used for concrete where both normal and shear stresses can determine failure. The Drucker-Prager yield criterion may be expressed as

$$\mathcal{F}^{\mathrm{DP}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \alpha \mathcal{I}_1(\boldsymbol{\sigma}) + k = 0 \tag{5.52}$$

it should probably be -k .. this needs to be propagated below!

　　The problem is then to choose α and k in a meaningful way, typically in relation with the Mohr-Coulomb criterion.
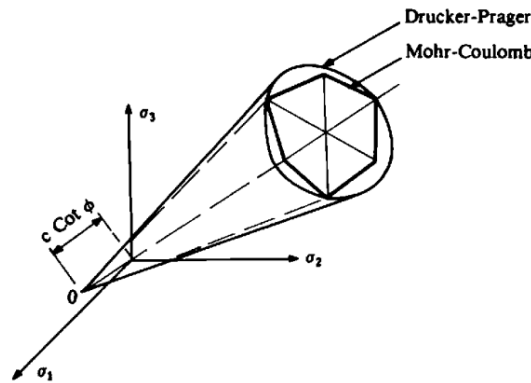
(a) **Geometrical representation of the Mohr–Coulomb and Drucker–Prager yield surfaces in principal stress space.**

Taken from Owen and Hinton [1026]. The Drucker-Prager yield envelope circumscribes the Mohr-Coulomb one.

　　Using the parameters σ_m, τ_m, $a = -\sqrt{3}\tan\theta$, $\mathcal{I}_1(\boldsymbol{\sigma})$ and $\mathcal{I}_2(\boldsymbol{\tau})$ of Section 4.6 we have

$$
\begin{aligned}
\mathcal{F}^{\mathrm{DP}} &= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \alpha\mathcal{I}_1(\boldsymbol{\sigma}) + k \\
&= \sqrt{\frac{\tau_m^2}{3}(a^2+3)} + \alpha(3\sigma_m - a\tau_m) + k \\
&= \tau_m\sqrt{(a^2/3+1)} + \alpha(3\sigma_m + \tau_m\sqrt{3}\tan\theta) + k \qquad (\text{since } \tau_m > 0) \\
&= \tau_m\sqrt{\tan^2\theta + 1} + \alpha(3\sigma_m + \tau_m\sqrt{3}\tan\theta) + k \\
&= \tau_m\sqrt{\frac{1}{\cos^2\theta}} + \alpha(3\sigma_m + \tau_m\sqrt{3}\tan\theta) + k \\
&= \tau_m\frac{1}{\cos\theta} + \alpha(3\sigma_m + \tau_m\sqrt{3}\tan\theta) + k \qquad (\text{since } \cos\theta > 0)
\end{aligned}
$$

$\mathcal{F}^{\mathrm{DP}} = 0$ then leads to write

$$
\begin{aligned}
\tau_m + (3\alpha\sigma_m + k)\cos\theta + \tau_m\alpha\sqrt{3}\sin\theta &= 0 \\
\Rightarrow \quad \tau_m(1 + \alpha\sqrt{3}\sin\theta) + (3\alpha\sigma_m + k)\cos\theta &= 0
\end{aligned}
$$

and finally

$$\tau_m = -\frac{(3\alpha\sigma_m + k)\cos\theta}{1 + \alpha\sqrt{3}\sin\theta} = -\frac{3\alpha\cos\theta}{1 + \alpha\sqrt{3}\sin\theta}\sigma_m - \frac{k\cos\theta}{1 + \alpha\sqrt{3}\sin\theta}$$

Remark. *This is the same equation as Eq. 19 of Wojciechowski [1467] but with $\theta \to -\theta$.*

The Mohr-Coulomb yield criterion is (see Eq. (5.42))

$$\tau_m = -\sigma_m \sin\phi + c\cos\phi$$

so that equating both expressions of τ_m for the Drucker-Prager and Mohr-Coulomb criteria leads to:

$$-\frac{3\alpha\cos\theta}{1+\alpha\sqrt{3}\sin\theta} = -\sin\phi \tag{5.53}$$

$$-\frac{k\cos\theta}{1+\alpha\sqrt{3}\sin\theta} = c\cos\phi \tag{5.54}$$

Eq. (5.53) yields

$$3\alpha\cos\theta = \sin\phi(1+\alpha\sqrt{3}\sin\theta)$$
$$\Rightarrow \quad 3\alpha\cos\theta - \alpha\sqrt{3}\sin\theta\sin\phi = \sin\phi$$

and finally

$$\boxed{\alpha(\phi) = \frac{\sin\phi}{3\cos\theta - \sqrt{3}\sin\theta\sin\phi}} \tag{5.55}$$

Inserting this into Eq. (5.54):

$$
\begin{aligned}
-k\cos\theta &= c\cos\phi\left(1+\alpha\sqrt{3}\sin\theta\right) \\
&= c\cos\phi\left(1+\frac{\sin\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\sqrt{3}\sin\theta\right) \\
&= c\cos\phi\left(1+\frac{\sqrt{3}\sin\phi\sin\theta}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\right) \\
&= c\cos\phi\left(\frac{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}+\frac{\sqrt{3}\sin\phi\sin\theta}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\right) \\
&= c\cos\phi\left(\frac{3\cos\theta}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\right)
\end{aligned}
$$

so that

$$\boxed{k(c,\phi) = -\frac{3\,c\cos\phi}{3\cos\theta - \sqrt{3}\sin\theta\sin\phi}} \tag{5.56}$$

The Drucker-Prager yield criterion which for a given θ is equal to the Mohr-Coulomb yield is then:

$$
\begin{aligned}
\mathcal{F}^{\text{DP}} &= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \alpha(\phi)\mathcal{I}_1(\boldsymbol{\sigma}) + k(c,\phi) \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{\sin\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\mathcal{I}_1(\boldsymbol{\sigma}) - \frac{3\,c\cos\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi} \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{3\sin\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{3\,c\cos\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[\frac{3\,p\sin\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi} + \frac{3\,c\cos\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi}\right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{3\,p\sin\phi + 3\,c\cos\phi}{3\cos\theta-\sqrt{3}\sin\theta\sin\phi} \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{p\sin\phi + c\cos\phi}{\cos\theta - \frac{1}{\sqrt{3}}\sin\theta\sin\phi}
\end{aligned}
\tag{5.57}
$$

$$\tag{5.58}$$

which, when multiplied by $\cos\theta - \frac{1}{\sqrt{3}}\sin\theta\sin\phi$, gives the Mohr-Coulomb criterion of Eq. (5.44).

5.14.1 The three standard types

For $\theta = \pi/6$, the DP yield surface **circumscribes** the MC yield surface and Eq. (5.57) writes:

$$
\begin{aligned}
\mathcal{F}^{\mathrm{DP}} &= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{3\sin\phi}{3\sqrt{3}/2 - \sqrt{3}/2\,\sin\phi} \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{3\,c\cos\phi}{3\sqrt{3}/2 - \sqrt{3}/2\,\sin\phi} \right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{6\sin\phi}{\sqrt{3}(3 - \sin\phi)} \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{6\,c\cos\phi}{\sqrt{3}(3 - \sin\phi)} \right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{6p\sin\phi + 6\,c\cos\phi}{\sqrt{3}(3 - \sin\phi)}
\end{aligned}
\tag{5.59}
$$

i.e.
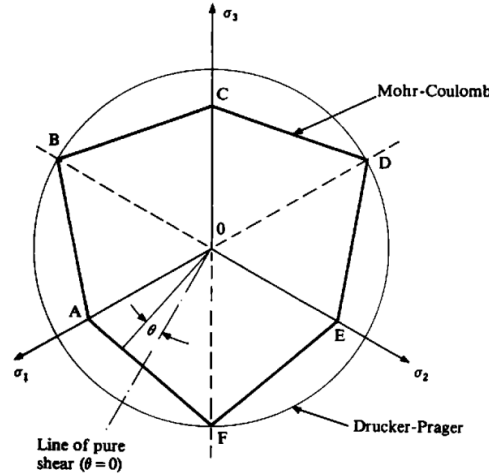
(b) Two-dimensional, π plane, representation of the Mohr–Coulomb and Drucker–Prager yield criteria.

Taken from Owen and Hinton [1026].

$$
\boxed{\mathcal{F}^{\mathrm{DP}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{6\sin\phi}{\sqrt{3}(3 - \sin\phi)} \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} - \frac{6\,c\cos\phi}{\sqrt{3}(3 - \sin\phi)}}
\tag{5.60}
$$

which is the formula used in Glerum *et al.* (2018) [496]. This is also Eq. (14a) in Zienkiewicz & Cormeau (1974) [1533], Eq. (7.18) in Owen and Hinton [1026], and Eq. (13.10a) in Zienkiewicz (1975) [1532] provided it is divided altogether by $\sqrt{3}$.

For $\theta = -\pi/6$, the DP yield surface **middle circumscribes** the MC yield surface and Eq. (5.57) writes:

$$
\begin{aligned}
\mathcal{F}^{\mathrm{DP}} &= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{3\sin\phi}{3\sqrt{3}/2 + \sqrt{3}/2\sin\phi} \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{3\,c\cos\phi}{3\sqrt{3}/2 + \sqrt{3}/2\sin\phi} \right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{6\sin\phi}{\sqrt{3}(3 + \sin\phi)} \frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{6\,c\cos\phi}{\sqrt{3}(3 + \sin\phi)} \right) \right] \\
&= \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{6p\sin\phi + 6c\cos\phi}{\sqrt{3}(3 + \sin\phi)}
\end{aligned}
\tag{5.61}
$$

This is Eq. (7.19) of Owen and Hinton [1026].

Another DP formulation which **inscribes** the MC yield surface is found on the wikipedia page of the Drucker-Prager yield criterion [10]:

---

[10] `https://en.wikipedia.org/wiki/Drucker-Prager_yield_criterion`. (but I have no idea how it is arrived at)

The Drucker–Prager yield criterion has the form

$$\sqrt{J_2} = A + B\,I_1$$

where I_1 is the first invariant of the Cauchy stress and J_2 is the second invariant of the deviatoric part of the Cauchy stress. The constants A, B are determined from experiments.

$$A = \frac{6\,c\,\cos\phi}{\sqrt{3}(3 - \sin\phi)} \quad ; \quad B = \frac{2\,\sin\phi}{\sqrt{3}(3 - \sin\phi)}$$

If the Drucker–Prager yield surface **middle circumscribes** the Mohr–Coulomb yield surface then

$$A = \frac{6\,c\,\cos\phi}{\sqrt{3}(3 + \sin\phi)} \quad ; \quad B = \frac{2\,\sin\phi}{\sqrt{3}(3 + \sin\phi)}$$

If the Drucker–Prager yield surface **inscribes** the Mohr–Coulomb yield surface then

$$A = \frac{3\,c\,\cos\phi}{\sqrt{9 + 3\,\sin^2\phi}} \quad ; \quad B = \frac{\sin\phi}{\sqrt{9 + 3\,\sin^2\phi}}$$

With our notations:

$$\mathcal{F}^{\mathrm{DP}} \;=\; \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \left[-\frac{3\sin\phi}{\sqrt{9 + 3\sin^2\phi}}\frac{\mathcal{I}_1(\boldsymbol{\sigma})}{3} + \frac{3\,c\cos\phi}{\sqrt{9 + 3\sin^2\phi}} \right] \tag{5.62}$$

The yield surfaces of these three Drucker-Prager formulations are plotted against the Mohr-Coulomb yield surface in Section 5.17.

Remark. *Leroy and Ortiz [818] (1989) use the Drucker-Prager plasticity model also and match it to the Mohr-Coulomb model in the triaxial test and formulate it as follows (Their definition of the second invariant of stress contains a 3/2 term):*

$$
\begin{aligned}
\mathcal{F} \;&=\; \tau_e\sqrt{3} + \frac{6\sin\phi}{3 - \sin\phi}\left(-p - \frac{c}{\tan\phi} \right) \\
&=\; \tau_e\sqrt{3} - \left(\frac{6\sin\phi}{3 - \sin\phi}p + c\frac{6\cos\phi}{3 - \sin\phi} \right) \\
&=\; \sqrt{3}\left[\tau_e - \left(\frac{6\sin\phi}{\sqrt{3}(3 - \sin\phi)}p + c\frac{6\cos\phi}{\sqrt{3}(3 - \sin\phi)} \right) \right]
\end{aligned}
\tag{5.63}
$$

Except for the $\sqrt{3}$ this is identical to Eq. (5.59).

Remark. *Bui et al. (2008) [174] use yet again another formulation:*

$$
\begin{aligned}
\mathcal{F} \;&=\; \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{\tan\phi}{\sqrt{9 + 12\tan^2\phi}}\mathcal{I}_1(\boldsymbol{\sigma}) - \frac{3c}{\sqrt{9 + 12\tan^2\phi}} \\
&=\; \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{\sin\phi}{\sqrt{9\cos^2\phi + 12\sin^2\phi}}\mathcal{I}_1(\boldsymbol{\sigma}) - \frac{3c\,\cos\phi}{\sqrt{9\cos^2\phi + 12\sin^2\phi}} \\
&=\; \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{\sin\phi}{\sqrt{9 + 3\sin^2\phi}}\mathcal{I}_1(\boldsymbol{\sigma}) - \frac{3c\,\cos\phi}{\sqrt{9 + 3\sin^2\phi}}
\end{aligned}
$$

which is identical to (5.62).

Remark. *Cacace & Jacquey (2017) [212] replace $\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$ by $\sqrt{\mathcal{I}_2(\boldsymbol{\tau}) + \epsilon_0^2}$ where ϵ_0 is a small non-hardening parameters here introduced to relax the singularity at the cone's tip of the Drucker-Prager yield envelope.*

📚 Relevant Literature: Gilda Currenti and Charles A Williams. "Numerical modeling of deformation and stress fields around a magma chamber: Constraints on failure conditions and rheology". In: *Phys. Earth. Planet. Inter.* 226 (2014), pp. 14–27. DOI: 10.1016/j.pepi.2013.11.003

5.14.2 What happens when $\phi = 0$?

From (5.55) we obtain $\alpha = 0$, and from Eq. (5.56) we obtain

$$k = -\frac{3c}{3\cos\theta_{\mathrm{L}}}$$

Then

$$\mathcal{F}^{\mathrm{DP}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{c}{\cos\theta_{\mathrm{L}}} = 0 \tag{5.64}$$

The MC-circumscribing DP formulation ($\theta_L = \pi/6$) then becomes

$$\mathcal{F}^{\mathrm{DP}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{6\,c}{3\sqrt{3}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{2c}{\sqrt{3}} \tag{5.65}$$

We obviously obtain an identical expression for the middle-circumscribing formulation ($\theta_L = -\pi/6$). The MC-inscribing case is then

$$\mathcal{F}^{\mathrm{DP}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{3\,c\cos\phi}{\sqrt{9}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - c$$

reconcile with von Mises!

5.14.3 Dissecting the original paper by Drucker and Prager (1952)

The authors state that a yield function which is a proper generalisation of the M-C hypothesis is:

$$\mathcal{F} = \alpha\mathcal{I}_1(\boldsymbol{\sigma}) + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k$$

where α and k are positive constants at each point of the material.

According to the concept of plastic potential, the stress-train relation corresponding to this yield function is

$$\dot{\varepsilon}_{ij}^{p} = \lambda\frac{\partial\mathcal{F}}{\partial\sigma_{ij}}$$

where $\dot{\varepsilon}_{ij}^{p}$ is the plastic strain rate and λ is a positive factor of proportionality which may assume different values in space. Using the above expression for $\mathcal{F}$:

$$\dot{\varepsilon}_{ij}^{p} = \lambda\left(\alpha\delta_{ij} + \frac{\tau_{ij}}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\right) \tag{5.66}$$

A very important feature of this equation is that the plastic rate of cubical dilation is

$$\mathrm{tr}[\dot{\boldsymbol{\varepsilon}}^{p}] = \dot{\varepsilon}_{ii}^{p} = 3\alpha\lambda \neq 0 \tag{5.67}$$

This equation shows that plastic deformation must be accompanied by an increase in volume if $\alpha \neq 0$. This property is known as dilatancy.

Plane strain   We need to establish three expressions. First, from Eq. (5.66) we can write

$$\dot{\varepsilon}_{zz}^{p} = \lambda\left(\alpha + \frac{\tau_{zz}}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\right)$$

but since $\dot{\varepsilon}_{zz} = 0$ in plane strain then we find

$$\tau_{zz} = -2\alpha\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{5.68}$$

which is Eq. (6) of the paper.

Second, we start from the definition of the first invariant and use the equation above:

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_{xx} + \sigma_{yy} + \sigma_{zz}$$

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_{xx} + \sigma_{yy} + \tau_{zz} + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})$$

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_{xx} + \sigma_{yy} - 2\alpha\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})$$

$$\frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma}) = \sigma_{xx} + \sigma_{yy} - 2\alpha\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$$

$$\mathcal{I}_1(\boldsymbol{\sigma}) = \frac{3}{2}(\sigma_{xx} + \sigma_{yy}) - 3\alpha\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{5.69}$$

which is Eq. (7) of the paper.

Finally, we start from (and we use the fact that $\sigma_{xx} - \sigma_{yy} = \tau_{xx} - \tau_{yy}$)

$$\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \tau_{xy}^2 = \frac{1}{4}(\sigma_{xx} - \sigma_{yy})^2 + \tau_{xy}^2$$

$$= \frac{1}{4}(\sigma_{xx} - \sigma_{yy})^2 + \underbrace{\tau_{xy}^2 + \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2)}_{\mathcal{I}_2(\boldsymbol{\tau})} - \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2)$$

$$= \frac{1}{4}(\tau_{xx} - \tau_{yy})^2 + \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{2}\tau_{xx}^2 - \frac{1}{2}\tau_{yy}^2 - \frac{1}{2}\tau_{zz}^2$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) + \frac{1}{4}\tau_{xx}^2 - \frac{1}{2}\tau_{xx}\tau_{yy} + \frac{1}{4}\tau_{yy}^2 - \frac{1}{2}\tau_{xx}^2 - \frac{1}{2}\tau_{yy}^2 - \frac{1}{2}4\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{4}\tau_{xx}^2 - \frac{1}{2}\tau_{xx}\tau_{yy} - \frac{1}{4}\tau_{yy}^2 - 2\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{4}(\tau_{xx}^2 + 2\tau_{xx}\tau_{yy} + \tau_{yy}^2) - 2\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{4}\underbrace{(\tau_{xx} + \tau_{yy})^2}_{-\tau_{zz}} - 2\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{4}\tau_{zz}^2 - 2\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{4}4\alpha^2\mathcal{I}_2(\boldsymbol{\tau}) - 2\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau}) - 3\alpha^2\mathcal{I}_2(\boldsymbol{\tau})$$

$$= \mathcal{I}_2(\boldsymbol{\tau})(1 - 3\alpha^2) \tag{5.70}$$

so that

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{1 - 3\alpha^2}\left[\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \tau_{xy}^2\right] \tag{5.71}$$

which is Eq. (8) of the paper.

In the paper the authors propose the yield function

$$\mathcal{F} = \alpha\mathcal{I}_1(\boldsymbol{\sigma}) + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k$$

We first replace the first (plane strain) invariant (see Eq. (5.69)):

$$\mathcal{F} = \alpha\left[\frac{3}{2}(\sigma_{xx} + \sigma_{yy}) - 3\alpha\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\right] + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k$$

$$= \alpha\frac{3}{2}(\sigma_{xx} + \sigma_{yy}) - 3\alpha^2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k$$

$$= \alpha\frac{3}{2}(\sigma_{xx} + \sigma_{yy}) + (1 - 3\alpha^2)\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k$$

and we now introduce the second invariant of Eq. (5.71):

$$
\begin{aligned}
\mathcal{F} &= \alpha\frac{3}{2}(\sigma_{xx}+\sigma_{yy}) + (1-3\alpha^2)\frac{1}{(1-3\alpha^2)^{1/2}}\left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - k \\
&= \alpha\frac{3}{2}(\sigma_{xx}+\sigma_{yy}) + (1-3\alpha^2)^{1/2}\left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - k \\
&= \frac{3\alpha}{(1-3\alpha^2)^{1/2}}\frac{1}{2}(\sigma_{xx}+\sigma_{yy}) + \left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - \frac{k}{(1-3\alpha^2)^{1/2}} \\
&= \underbrace{\frac{3\alpha}{(1-3\alpha^2)^{1/2}}}_{\sin\phi}\frac{1}{2}(\sigma_{xx}+\sigma_{yy}) + \left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - \underbrace{\frac{k}{(1-12\alpha^2)^{1/2}}}_{c}\underbrace{\frac{(1-12\alpha^2)^{1/2}}{(1-3\alpha^2)^{1/2}}}_{\cos\phi}
\end{aligned}
\tag{5.72}
$$

Note that if we define a triangle with sides $(1-12\alpha^2)^{1/2}$ and 3α with hypotenuse $(1-3\alpha^2)^{1/2}$ then the angle ϕ makes sense and we recover $\cos^2\phi + \sin^2\phi = 1$.

In the end:

$$
\mathcal{F} = \left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - \left(-\frac{1}{2}(\sigma_{xx}+\sigma_{yy})\sin\phi + c\cos\phi\right)
$$

which is the Mohr-Coulomb yield criterion of Eq. (1) in the paper.

Note that when $\alpha = 0$ (yield criterion independent of the mean stress - incompressible flow see Eq. (5.67)) then $c = k$, $\cos\phi = 1$ and $\sin\phi = 0$ and we find the Tresca yield criterion

$$
F^{TR} = \left[\left(\frac{\sigma_{xx}-\sigma_{yy}}{2}\right)^2+\tau_{xy}^2\right]^{1/2} - k
$$

Also, setting $\alpha = 0$ in Eq. (5.71) yields a criterion that writes

$$
F^{vM} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - k
$$

which is the von Mises criterion!

We now look into the derivatives of the Drucker-Prager plastic potential $\mathcal{Q}^{\text{DP}}(\boldsymbol{\sigma})$. We have

$$
\mathcal{Q}^{\text{DP}}(\boldsymbol{\sigma}) = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + \alpha\mathcal{I}_1(\boldsymbol{\sigma}) + k
$$

Then

$$
\frac{\partial\mathcal{Q}^{\text{DP}}}{\partial\mathcal{I}_1(\boldsymbol{\sigma})} = \alpha \tag{5.73}
$$

$$
\frac{\partial\mathcal{Q}^{\text{DP}}}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} = 1 \tag{5.74}
$$

$$
\frac{\partial\mathcal{Q}^{\text{DP}}}{\partial\theta_{\text{L}}(\boldsymbol{\tau})} = 0 \tag{5.75}
$$

The parameters α and k can be expressed as a function of the angle of friction and cohesion so as to match the Mohr-Coulomb criterion in some sense (see above). Then

$$
C_1^{\text{DP}} = \alpha \tag{5.76}
$$

$$
C_2^{\text{DP}} = \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \tag{5.77}
$$

$$
C_3^{\text{DP}} = 0 \tag{5.78}
$$

5.15   The Griffith-Murrell failure criterion

The Griffith-Murrell yield criterion is not often used. Noticeable exceptions are Braun [148] (1994), Braun and Beaumont [151] (1995), Batt and Braun [58] (1997). In Braun [148] (1994) we read:

> Extending the work of Griffith (1921) to three dimensional stress distributions, Murrell (1963) suggested the following criterion for rock failure expressed in terms of the principal stresses:
> $$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 + 24 T_0 (\sigma_1 + \sigma_2 + \sigma_3) = 0$$
> where T_0 is a material property called the tensile strength. In principal stress space, this criterion is represented by a paraboloid of revolution around the pressure (or hydrostatic) axis.

Using the definition of $\mathcal{I}_2(\boldsymbol{\tau})$ of Eq. (4.75) and $\mathcal{I}_1(\boldsymbol{\sigma})$ of Eq. (4.74), it also writes:

$$6 \left[\underbrace{\frac{1}{6}(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2}_{\mathcal{I}_2(\boldsymbol{\tau})} \right] + 24 T_0 \underbrace{(\sigma_1 + \sigma_2 + \sigma_3)}_{\mathcal{I}_1(\boldsymbol{\sigma})} = 0$$

or,

$$\mathcal{I}_2(\boldsymbol{\tau}) + 4 T_0 \mathcal{I}_1(\boldsymbol{\sigma}) = 0$$

which is the formulation found in Braun and Beaumont [151] (1995). Furthermore the authors add:

> A material that obeys this failure criterion can be interpreted to be a frictional material in which the strength increases in proportion to the square root of the pressure. The failure envelope, $\mathcal{F} = 0$, is a paraboloid of revolution about the pressure axis in principal stress space in contrast to the prismatic cone of the Mohr-Coulomb criterion.

It is also worth mentioning that the authors explain the numerical methods (FEM, plasticity implementation, solver) in the appendix.

Note that Hansen, Nielsen, and Lykke-Andersen [566] (2000) use an identical formulation, although the authors use the lithostatic pressure instead of the full pressure. They also use a tensile strength parameter T_0^e and a compressive strength parameter T_0^c, both around a few tens of MPas. They refer to the book by Jaeger & Cook (1979) when presenting the Griffith criterion. We find there (new edition of 2007, chapter 10.9) the same equation but with a minus sign:

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = 24 T_0 (\sigma_1 + \sigma_2 + \sigma_3)$$

It is not easy to find numerical values for the T_0 value, but in Batt and Braun [58] (1997) we find in Table 2: $T_0 = 10^7$ Pa and in Hansen, Nielsen, and Lykke-Andersen [566] we find this great useful table:

Symbol	Meaning	Mantle Wet Olivine	Lower crust Wet Feldspar	Upper crust Wet Quartz	Sediments Wet Quartz
E	Young's modulus	10^{11} Pa	10^{11} Pa	10^{11} Pa	10^{11} Pa
ν	Poisson' ratio	0.25	0.25	0.25	0.25
n	Creep parameter	4.48	3.20	3.10	3.10
B^*	Creep parameter	0.2628 MPa s$^{1/4.48}$	12.28 MPa s$^{1/3.20}$	208.0 MPa s$^{1/3.10}$	208.0 MPa s$^{1/3.10}$
Q^*	Creep parameter	498 kJ/mol	239 kJ/mol	135 kJ/mol	135 kJ/mol
T_0^{c}	Tensile strength parameter	26.2 MPa	13.1 MPa	13.1 MPa	8.73 MPa
T_0^{c}	Compressive strength parameter	52.4 MPa	26.2 MPa	26.2 MPa	17.5 MPa
ρ	Density	3300 kg/m^3	2900 kg/m^3	2700 kg/m^3	2300 kg/m^3
k	Thermal conductivity	4.0 W/m/K	2.3 W/m/K	3.0 W/m/K	2.0 W/m/K
c	Specific heat	1000 W/kg/K	900.0 W/kg/K	850.0 W/kg/K	1000.0 W/kg/K
A	Radiogenic heat production	0.01 μW/m^3	0.3 μW/m^3	1.3 μW/m^3	1.3 μW/m^3

5.16 The Cam-clay failure criterion

The Original Cam-Clay model is based on the assumption that the soil is isotropic, elasto-plastic, deforms as a continuum, and it is not affected by creep.

♻ Relevant Literature: [1052]

ask Chris Spiers. Pijnenburg *et al.* , JGR 2019

5.17 The failure envelope, or yield surface

♻ Relevant Literature: Schöpfer *et al.* (2013) [1212].

A yield surface is a five-dimensional surface in the six-dimensional space of stresses. The state of stress of inside the yield surface is elastic. When the stress state lies on the surface the material is said to have reached its yield point and the material is said to have become plastic. Further deformation of the material causes the stress state to remain on the yield surface, even though the surface itself may change shape and size as the plastic deformation evolves, this is because stress states that lie outside the yield surface are non-permissible.

The yield surface is usually expressed in terms of (and visualized in) a three-dimensional principal stress space $(\sigma_1, \sigma_2, \sigma_3)$, a two- or three-dimensional space spanned by stress invariants or a version of the three-dimensional Haigh-Westergaard space[11].
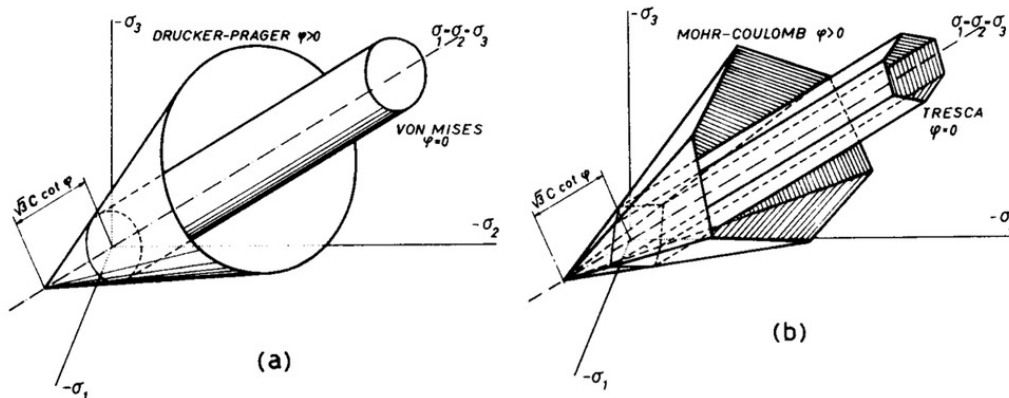
Yield surfaces in stress space [1533]. Note that the axes are $-\sigma_1, -\sigma_2, -\sigma_3$

Having obtained the equations for the yield functions in the previous sections, we can easily test them as follows: in the $(\sigma_1, \sigma_2, \sigma_3)$ space we can look for stress states that fulfil the yield equations.

---

[11]https://en.wikipedia.org/wiki/Stress_space

I set $c = 20$ MPa and $\phi = 20°$ and restrain the search to the space $[-100 \text{ MPa:} 100 \text{ MPa}]^3$. The python code and the gnuplot script used to generate the plots hereafter are in `images/rheology/surfaces`. The implemented algorithm is somewhat naive and quite inefficient: discretise the space in N^3 points and for each point check whether any of the von Mises, Tresca, Mohr-Coulomb and (the three variants of) Drucker-Prager criteria is satisfied and when the point is in the space $\sigma_1 + \sigma_2 + \sigma_3 = 10$MPa (perpendicular to the $x = y = z$ line) write it to the corresponding file.

The recovered surfaces are similar to those of the figure above but their plot in a 3D space is difficult. I have therefore isolated two sub-plots. The first one is for $\sigma_1 = \sigma_2$:
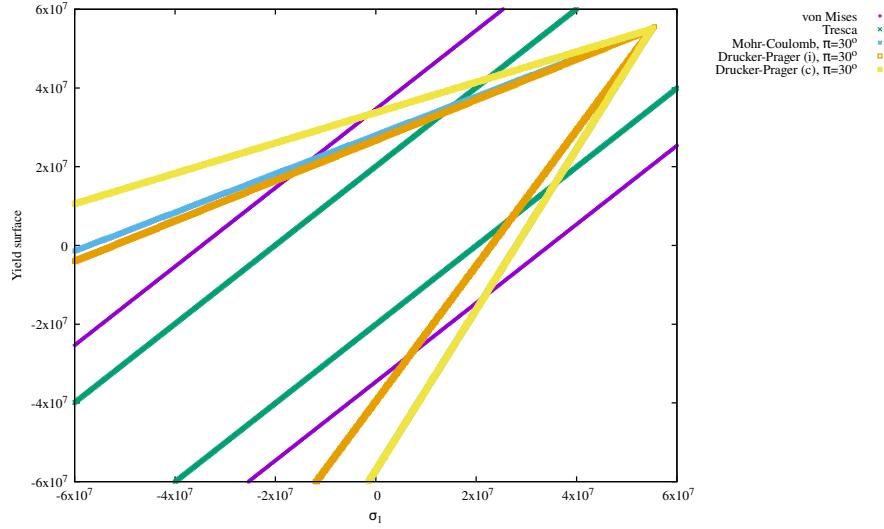
We see that the von Mises and Tresca envelopes are parallel to the line $\sigma_1 = \sigma_2 = \sigma_3$ (which is expected since they do not depend on pressure).

The second plot is in the plane $\sigma_1 + \sigma_2 + \sigma_3 = 0$ which is perpendicular to the middle line $\sigma_1 = \sigma_2 = \sigma_3 = 0$. To facilitate plotting the envelopes are plotted as a function of σ_1 only (so that even though they are circles in the chosen plane they appear here as ellipses):
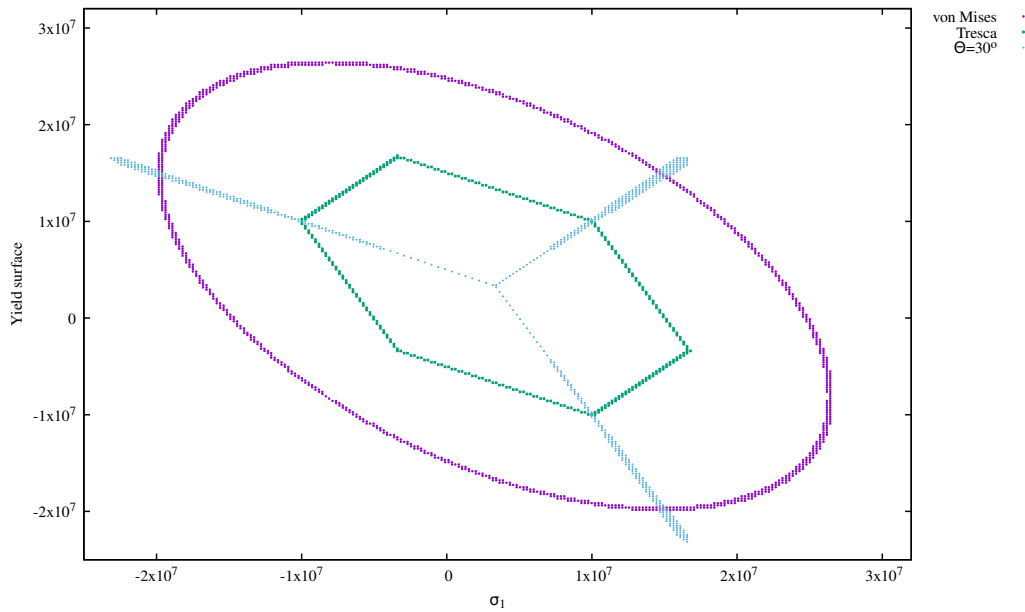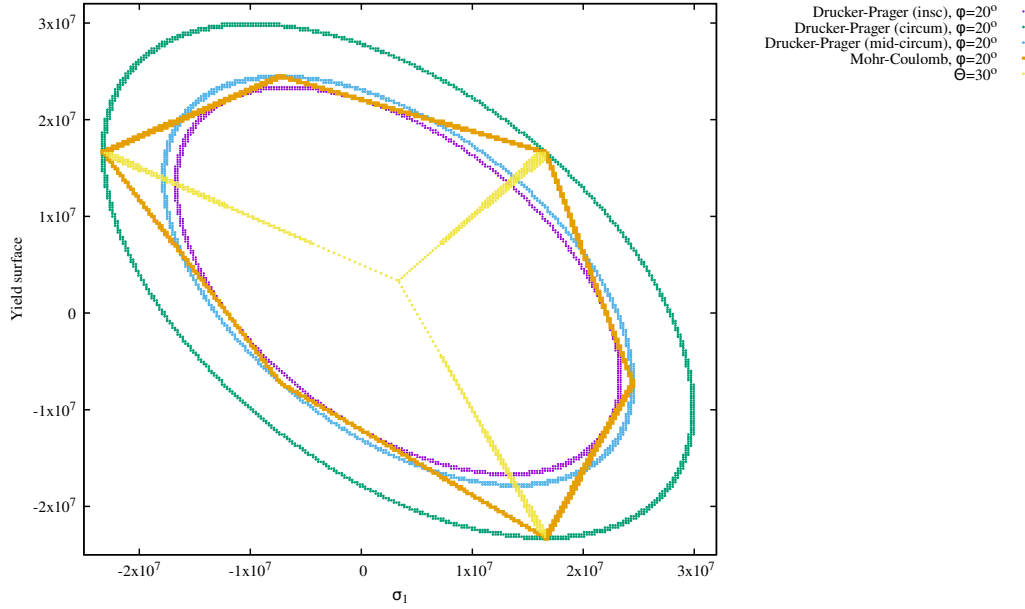
We see that we indeed recover that the three Drucker-Prager formulations inscribe (purple), middle-circumscribe (blue) and circumscribe (green) the Mohr-Coulomb one.

the code for these plots has been moved to stone 159. Update necessary, etc ...

5.18 Peierls creep

Looking at the literature, there seem to be many formulations for the Peierls creep deformation mechanism, but it it appears that a standard formulation for the Peierls creep writes:

$$\dot{\varepsilon} = A\sigma^n \exp\left[-\frac{Q+pV}{RT}\left(1-(\frac{\sigma}{\sigma_P})^k\right)^q\right]$$

and it seems common to take $k = 1$, and $n = 2$ [483, 716]

$$\dot{\varepsilon} = A\sigma^2 \exp\left[-\frac{Q+pV}{RT}\left(1-\frac{\sigma}{\sigma_P}\right)^q\right]$$

Elbeshausen & Melosh (2018) [389] use

$$\dot{\varepsilon} = A \exp\left[-\frac{Q}{RT}\left(1-\frac{\sigma}{\sigma_P}\right)^q\right]$$

In Chenin *et al.* (2019) [247] the authors state that their Peierls creep implementation relies on parameters from Evans and Goetze (1979) [408] using the approach of Kameyama *et al.* (1999) [704]:

$$\eta^{pe} = \frac{2}{3}\frac{(1-s)/s}{(1+s)/2s} A \left(\varepsilon_e^{ds}\right)^{\frac{1}{n}-1}$$

with A for this formulation:

$$A = \left[A_p \exp\left(-\frac{Q(1-\gamma)^2}{RT}\right)\right]^{-1/s} \gamma\sigma_p$$

where s is an effective stress exponent that depends on the temperature:

$$s = 2\gamma \frac{Q}{RT}(1 - \gamma)$$

where γ is a fitting parameter.

📚 Relevant LiteratureBabeyko, Sobolev, Vietor, Oncken, and Trumbull [39], Burov [201], Faul, Gerald, Farlai, Ahlefeldt, and Jackson [413], Garel, Goes, Davies, Davies, Kramer, and Wilson [461], Gerya [483], Goetze and Evans [499], Katayama and Karato [716], Kawazoe, Karato, Otsuka, Jing, and Mookherjee [726], Karato, Riedel, and Yuen [712], Mei, Suzuki, Kohlstedt, Dixon, and Durham [909], Zhong and Watts [1522], Chenin, Schmalholz, Manatschal, and Karner [246], Shi, Wei, Li, Liu, and Liu [1238], Review article from 1966: Guyot & Dorn [548]

5.19 Stress limiting (pseudo-plastic) rheology

Taken from van Hunen *et al.* (2002) [1403]:

$$\eta_y = \tau_y \dot{\varepsilon}_y^{-1/n_y} \dot{\varepsilon}_e^{(1/n_y)-1}$$

where the yield stress τ_y, the yield strain rate $\dot{\varepsilon}_y$ and the yield exponent n_y are prescribed parameters. In this article, $n_y = 10$, $\dot{\varepsilon}_y = 10^{-15}\text{s}^{-1}$ When $n_y = 1$ the viscosity is constant and given by $\eta_{eff} = \tau_y/\dot{\epsilon}_y$.

This rheology has also been coined pseudo-plastic in Zhong *et al.* (1998) [1520]. Their equation is simply

$$\eta_{eff} = A^{1/n} \dot{\varepsilon}_e^{-1+1/n}$$

where A is the preexponent which depends on temperature, pressure, and composition.
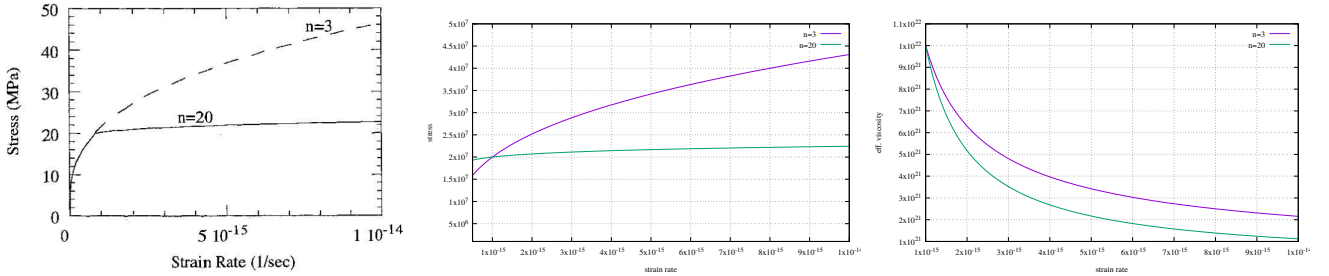
Left figure is taken from [1520]. Authors report $A = 7.9 \cdot 10^{-8}\text{Pa}^3\text{s}$ for the $n = 3$ case, which makes no sense. See gnuplot script for actual values of A.

5.20 Arrhenius law

A purely temperature-dependent dimensional Arrhenius law that emulates the temperature dependence of viscosity in silicate rock is often employed for mantle rocks [5, 1507, 1402, 134, 978, 1277, 105, 533]:

$$\eta(T) = \eta_0 \exp\left(\frac{Q}{R}(\frac{1}{T} - \frac{1}{T_0})\right) \qquad \text{or} \qquad \eta(T) = \eta_0 \exp\left(\frac{Q}{RT}\right) \tag{5.79}$$

where η_0 is a reference viscosity and T_0 its corresponding reference temperature.

It can also account for pressure effects as in [858] where the diffusion creep viscosity (under the assumption of homogeneous grain size) is temperature- and pressure-dependent:

$$\eta(T) = \eta_0 \exp\left(\frac{1}{R}(\frac{Q - pV}{T} - \frac{Q}{T_0})\right)$$

(I find the minus sign rather suspicious)

5.21   Frank-Kamenetskii rheology

Many CITCOMs-based publications [180, 179] have used the following (dimensionless) viscosity for the mantle:

$$\eta(T, z) = \eta_r(r) \exp(A(0.5 - T))$$

where η_r is a depth-dependent viscosity profile (usually defined as discontinuous linear profiles for various shells)

The non-dimensional activation coefficient is chosen to be $A = 9.2103$ in [179] which leads to a temperature-induced viscosity contrast of 10^4 (for $T \in [0, 1]$).

This is also called the Frank-Kamenetskii flow rule, as used in [1277, 801]:

$$\eta' = \eta_0 \exp(-\theta T)$$

where the the parameters η_0, θ account for the local chemical composition of the rock. Note that the Frank-Kamenetskii approximation takes many forms in the literature [1000].

Another temperature-dependent common expression is as follows [423]:

$$\eta(T) = \eta_\infty \exp\left(\frac{Q}{R}(\frac{1}{T} - \frac{1}{T_\infty})\right)$$

Also, following [423]: For studying transient convection in a non- Newtonian rheological fluid, it is expedient from a computational point of view to employ a law which behaves linearly for low stresses initially and becomes gradually non-Newtonian only after a certain threshold stress level has been surpassed [261, 263]:

$$\eta(T, p, \tau_2) = \eta(T, p) \frac{1}{A_2 + A_3 \tau_2^2}$$

where A_2 is a parameter describing the linear creep at low stress levels and A_3 governs the transition stress between Newtonian and non-Newtonian rheologies.

Coltice and Sheppard (2018) [293] use a depth- and temperature-dependent viscosity formulation:

$$\eta(z, T) = \eta_0(z) \exp \frac{Q}{RT}$$

Note that this expression is supplemented with a pseudo-plastic formulation [1154].

📥 Relevant Literature: [747]

5.22   Glen's law for ice

Ice and rocks share similarities in terms of (viscous) rheology. Glen's law is the most commonly used flow law for ice in glaciers and ice sheets [495] and it is actually a power-law type rheology:

$$\dot{\varepsilon} = A\boldsymbol{\tau}^n$$

with $n \sim 3$ and $A \sim 2.4 \cdot 10^{-24} \mathrm{Pa}^{-3} \cdot \mathrm{s}^{-1}$ at 0°C. The effective viscosity is then given by

$$\eta = \frac{1}{2A\tau_e^{n-1}}$$
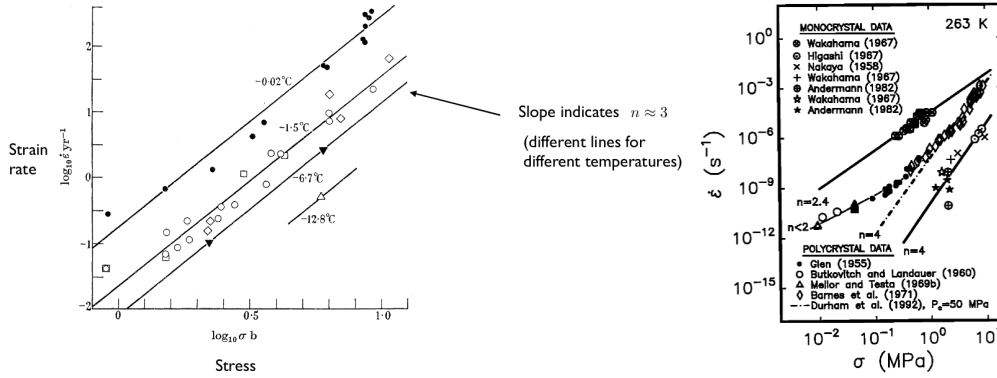
Left: Taken from Glen [495]; Right: taken from [500].

Most of these studies suggest values of the power-law exponent $n \sim 2 - 4$, and there seems to be a general indication that the exponent is lower at lower stresses.

The A coefficient above has been found to depend on temperature and is reasonably described with an Arrhenius law:

$$A(T) = A_0 \exp\left(-\frac{Q}{RT}\right)$$

A standard formulation is the Paterson-Budd law with a fixed Glen exponent $n = 3$ and a split Arrhenius term [1042]:

$$A = 3.615 \cdot 10^{-13} \mathrm{Pa}^{-3} \cdot \mathrm{s}^{-1}, \qquad Q = 60 \,\mathrm{kJ/mol}, \qquad if \quad T < 263\mathrm{K}$$

$$A = 1.733 \cdot 10^{3} \mathrm{Pa}^{-3} \cdot \mathrm{s}^{-1}, \qquad Q = 139 \,\mathrm{kJ/mol}, \qquad if \quad T > 263\mathrm{K}$$

Be careful that in these two equations the temperature T is the pressure-adjusted temperature [1042]. Note that A is also affected by the water content and the presence of impurities.

Finally, Glen's law is the standard rheology used for ice-sheet modelling but it does not account for the complex evolution of fabric and resulting anisotropy. Indeed the grain size evolution (growth & reduction) plays a large role in the rheology Behn, Goldsby, and Hirth [73] (2021). See STONE 59.

📥 Relevant LiteratureAlley [8] (1992), Greve [521] (1997), Greve and Blatter [522] (2009), Isaac, Stadler, and Ghattas [656] (2015), Krabbendam [772] (2016), Jiménez, Duddu, and Bassis [686] (2017), Helanow and Ahlkrona [594] (2018), Very mathematics heavy papers: [696, 240].

5.23 Strain rate partitioning across deformation mechanisms

When multiple viscous deformation mechanisms are present, one needs more dashpots, and more complicated element diagrams than the ones above occur (also when adding plastic deformation). Two important rules are to be remembered: 1) for parallel components, stresses are additive, strain rates are equal in each; 2) for components in series, stresses are equal in each and strain rates are additive. Let us then look in the next subsections at various assemblies of dashpots and plastic elements.

5.23.1 two viscous dampers in series
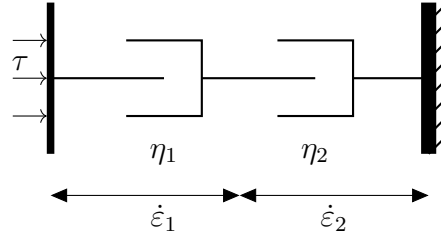
Each is subjected to the same stress τ but deforms with its own strain rate $\dot{\varepsilon}_1$ and $\dot{\varepsilon}_2$ and we have

$$\dot{\varepsilon}_T = \dot{\varepsilon}_1 + \dot{\varepsilon}_2 = \frac{\tau}{2\eta_1} + \frac{\tau}{2\eta_2} \tag{5.80}$$

The effective viscosity of this combination is denoted η_{eff} and is such that $\eta_{eff} = \tau/2\dot{\varepsilon}_T$, which means that

$$\frac{\tau}{2\eta_{eff}} = \frac{\tau}{2\eta_1} + \frac{\tau}{2\eta_2}$$

or,

$$\eta_{eff} = \left(\frac{1}{\eta_1} + \frac{1}{\eta_2} \right)^{-1}$$

i.e. it follows that the effective viscosity of two or more viscous dampers in series is the harmonic average of the individual viscosities of the dampers.

In general, for n dampers in series:

$$\eta_{eff} = \left(\sum_{i=1}^{n} \frac{1}{\eta_i} \right)^{-1}$$

5.23.2 two viscous dampers in parallel
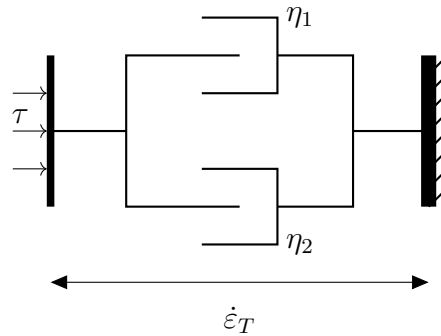
each is deformed with the same strain rate $\dot{\varepsilon}_T$ and their stresses add up:

$$\tau = \tau_1 + \tau_2 = 2\eta_1\dot{\varepsilon}_T + 2\eta_2\dot{\varepsilon}_T$$

and since we define the effective viscosity as $\tau = 2\eta_{eff}\dot{\varepsilon}_T$ then it follows:

$$2\eta_{eff}\dot{\varepsilon}_T = 2\eta_1\dot{\varepsilon}_T + 2\eta_2\dot{\varepsilon}_T$$

or,

$$\eta_{eff} = \eta_1 + \eta_2$$

i.e., the effective viscosity of two or more viscous dampers is the sum of their viscosities (*but not their arithmetic mean!*).

144

5.23.3 one viscous damper and a plastic element in parallel
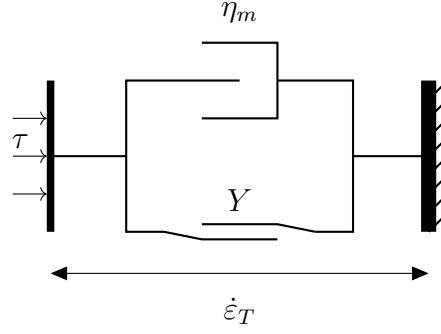
The effective 'plastic' viscosity of the plastic element is $\eta_p = \frac{Y}{2\dot{\varepsilon}_T}$ so the effective viscosity of this setup is then

$$\eta_{eff} = \frac{Y}{2\dot{\varepsilon}_T} + \eta_m$$

which is the viscosity of a Bingham fluid (see Section 5.5).

5.23.4 two viscous dampers and a plastic element
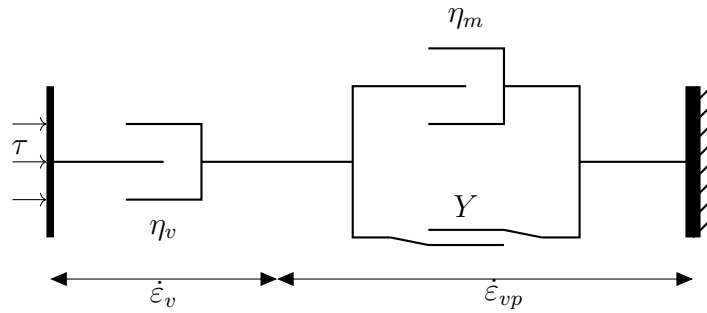
This rheology would be called visco-viscoplastic. The algorithm goes then as follows:

1. Assume we know η_v and $\dot{\varepsilon}_T$ (from previous iteration), as well as the plasticity parameters Y and η_m.

2. if $2\eta_v\dot{\varepsilon}_T < Y$ the stress is below the yield stress value and plasticity is not active. Use η_v in the material model and $\dot{\varepsilon}_v = \dot{\varepsilon}_T$.

3. if $2\eta_v\dot{\varepsilon}_T > Y$ the stress is above the yield value, which is not allowed. In this case the plastic element is 'switched on'. In that case the viscous damper is in series with the (visco)plastic element. The former deforms with a strain rate $\dot{\epsilon}_v$ while the latter with $\dot{\epsilon}_{vp}$ (both under the same stress τ) and we have $\dot{\varepsilon}_T = \dot{\varepsilon}_v + \dot{\varepsilon}_{vp}$.

$$\dot{\varepsilon}_T = \dot{\varepsilon}_v + \dot{\varepsilon}_{vp}$$

$$= \dot{\varepsilon}_v + \frac{\tau}{2\eta_{vp}}$$

$$= \dot{\varepsilon}_v + \frac{\tau}{2\left(\frac{Y}{2\dot{\varepsilon}_{vp}} + \eta_m\right)}$$

$$= \dot{\varepsilon}_v + \frac{\tau}{2\left(\frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)} + \eta_m\right)}$$

$$\dot{\varepsilon}_T - \dot{\varepsilon}_v = \frac{\tau}{2\left(\frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)} + \eta_m\right)}$$

$$2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)\left(\frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)} + \eta_m\right) = \tau$$

$$Y + 2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)\eta_m = \tau$$

$$Y + 2(\dot{\varepsilon}_T - \frac{\tau}{2\eta_v})\eta_m = \tau$$

$$Y + (2\eta_v\dot{\varepsilon}_T - \tau)\frac{\eta_m}{\eta_v} = \tau$$

$$Y + 2\eta_m\dot{\varepsilon}_T = \tau(1 + \frac{\eta_m}{\eta_v})$$

and finally

$$\tau = \frac{Y + 2\eta_m\dot{\varepsilon}_T}{1 + \frac{\eta_m}{\eta_v}} \tag{5.81}$$

Note that this solution exists even when $\eta_m = 0$, and then rather logically $\tau = Y$.

4. Once we have τ, we can easily compute $\dot{\epsilon}_v = \frac{\tau}{2\eta_v}$

5. We then compute $\dot{\varepsilon}_{vp} = \dot{\varepsilon}_T - \dot{\varepsilon}_v$ which we use to compute η_{vp}:

$$\eta_{vp} = \frac{Y}{2\dot{\varepsilon}_{vp}} + \eta_m$$

$$= \frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v)} + \eta_m$$

$$= \frac{Y}{2(\dot{\varepsilon}_T - \frac{\tau}{2\eta_v})} + \eta_m$$

$$= \frac{Y}{2(\dot{\varepsilon}_T - \frac{Y+2\eta_m\dot{\varepsilon}_T}{1+\frac{\eta_m}{\eta_v}}\frac{1}{2\eta_v})} + \eta_m$$

$$= \frac{Y}{2\dot{\varepsilon}_T - \frac{Y+2\eta_m\dot{\varepsilon}_T}{\eta_v+\eta_m}} + \eta_m \tag{5.82}$$

$$= \frac{Y(\eta_v + \eta_m)}{2(\eta_v + \eta_m)\dot{\varepsilon}_T - (Y + 2\eta_m\dot{\varepsilon}_T)} + \eta_m \tag{5.83}$$

$$= \frac{Y(\eta_v + \eta_m)}{2\eta_v\dot{\varepsilon}_T - Y} + \eta_m \tag{5.84}$$

$$= \frac{Y(\eta_v + \eta_m)/2\eta_v}{\dot{\varepsilon}_T - Y/2\eta_v} + \eta_m \tag{5.85}$$

6. Having obtained η_{vp} we can compute the final effective viscosity

$$\eta_{eff} = \left(\frac{1}{\eta_v} + \frac{1}{\eta_{vp}}\right)^{-1}$$

On the following plots are shown τ, $\dot{\varepsilon}_{vp}$, $\dot{\varepsilon}_v$, $\eta_v p$, and η_{eff} as a function of $\dot{\varepsilon}_T$:
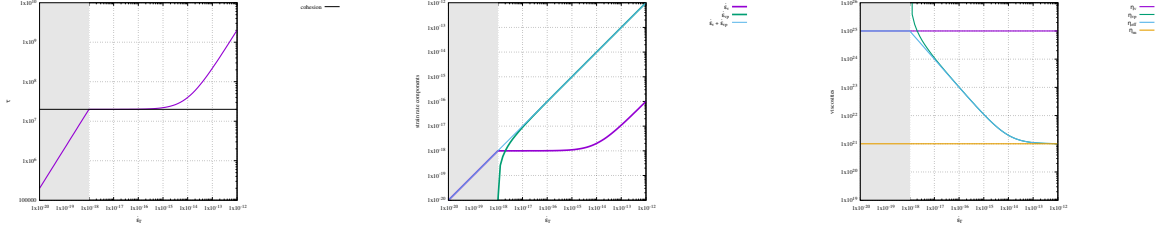
Obtained for $\eta_m = 10^{21}$, $Y = 20$MPa and $\eta_v = 10^{25}$. Python code in images/rheology/vvp/

In the following plots the resulting stress τ and effective viscosities η_{eff} are compared between the above approach ('new') and the simpler (and naive) approach where $\dot{\varepsilon}_T$ is used in η_{vp} instead of $\dot{\varepsilon}$ ('old'). In this particular case we see that it makes a difference at low strain rates close to the brittle-ductile transition.
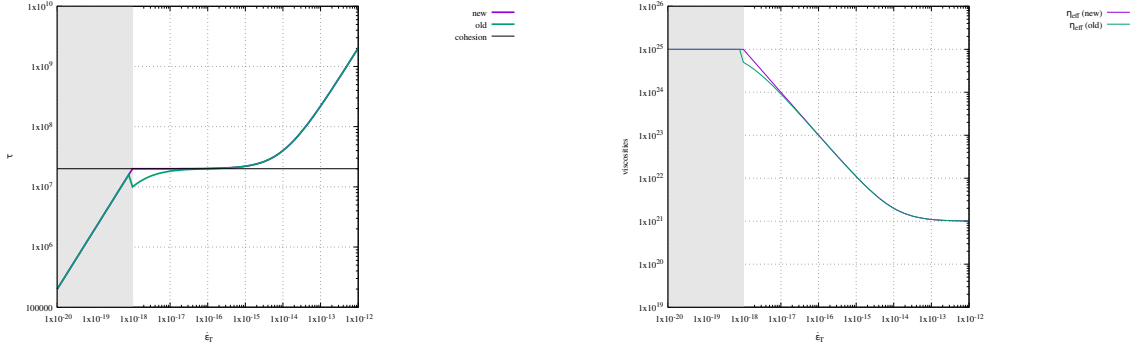
Obtained for $\eta_m = 10^{21}$, $Y = 20$MPa and $\eta_v = 10^{25}$. Python code in images/rheology/vvp/

Remark. *The introduction of the damper η_m in parallel with the plastic element has an unavoidable effect: the stress τ becomes larger than Y at high strain rate values! Since the vp block is akin to a bingham fluid, this is no surprise.*

Remark. *The viscous dashpot η_v also acts as a maximum viscosity cutoff: if η_{vp} becomes (very) large, i.e. $\eta_{vp} \gg \eta_v$, then $\eta_{eff} \to \eta_v$. Conversely, if $\eta_p = Y/2\dot{\varepsilon}_{vp}$ becomes (very) small, i.e. $\eta_p \ll \eta_m$ then η_m acts as a minimum viscosity limiter, i.e. $\eta_{vp} \to \eta_m$. Since $\eta_m \ll \eta_v$ then $\eta_{eff} \to \eta_m$.*

<u>A simple regularisation</u> This idea originates in Massmeyer *et al.* (2013) [889]. We postulate

$$\tilde{\eta}_{eff} = \left(1 - \exp(-\frac{\dot{\varepsilon}_T}{\dot{\varepsilon}_T^c})\right)\left(\frac{Y}{2\dot{\varepsilon}_T} + \eta_m\right)$$

where $\dot{\varepsilon}_T^c$ is the critical strain rate at which the transition viscous to viscous-viscoplastic occurs given by $\dot{\varepsilon}_T^c = Y/2\eta_v$. When $\dot{\varepsilon}_T \ll \dot{\varepsilon}_T^c$ then the exponential term tends to zero and

$$\tilde{\eta}_{eff} \to \frac{Y}{2\dot{\varepsilon}_T} + \eta_m$$

and if $\dot{\varepsilon}_T \to \infty$ then $\tilde{\eta}_{eff} \to \eta_m$. Conversely if $\dot{\varepsilon}_T \to 0$ then we can carry out a Taylor expansion of the exponential term ($\exp x \sim 1 + x$ when x is small).

$$\tilde{\eta}_{eff} \sim \left(\frac{\dot{\varepsilon}_T}{\dot{\varepsilon}_T^c}\right)\left(\frac{Y}{2\dot{\varepsilon}_T} + \eta_m\right) \to \frac{\dot{\varepsilon}_T}{\dot{\varepsilon}_T^c}\frac{Y}{2\dot{\varepsilon}_T} = \eta_v$$

At low strain rates the viscosity does not 'explode' but actually converges to the background viscosity η_v. The stress τ corresponding to this viscosity is simply $\tilde{\tau} = 2\tilde{\eta}_{eff}$. Both $\tilde{\tau}$ and $\tilde{\eta}_{eff}$ are plotted hereunder:
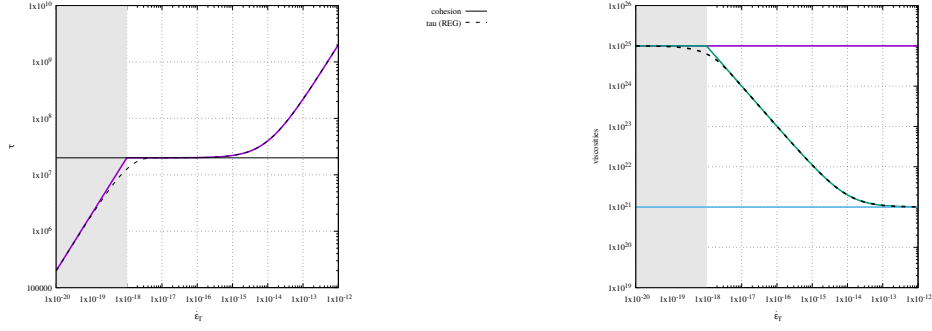
Obtained for $\eta_m = 10^{21}$, $Y = 20\text{MPa}$ and $\eta_v = 10^{25}$. Python code in images/rheology/vvp/
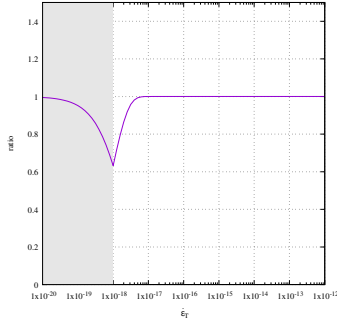
5.23.5 two nonlinear viscous dampers in series
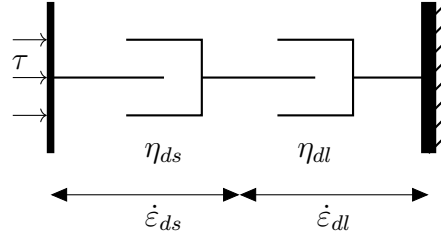
There are two dashpots in series, one accounts for dislocation creep, the other for diffusion creep. The algorithm goes then as follows:

1. Assume we know $\dot{\varepsilon}_T$ (from previous iteration).

2. The dashpots are in series so
$$\dot{\varepsilon}_T = \dot{\varepsilon}_{ds} + \dot{\varepsilon}_{df}$$

with

$$\dot{\varepsilon}_{ds} = A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) \tag{5.86}$$

$$\dot{\varepsilon}_{df} = A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) \tag{5.87}$$

such that we are in fact looking for the stress value τ so that

$$\dot{\varepsilon}_T = A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) + A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right)$$

148

or, we must find the zero of the function $\mathcal{F}(\tau)$:

$$\mathcal{F}(\tau) = \dot{\varepsilon}_T - A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) - A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right)$$

This equation can be solved with a Newton-Raphson algorithm and the iterations will be of the form:

$$\tau_{n+1} = \tau_n - \frac{\mathcal{F}(\tau_n)}{\mathcal{F}'(\tau_n)}$$

where the derivative of the function $\mathcal{F}$ with respect to τ reads:

$$\mathcal{F}'(\tau) = \frac{\partial \mathcal{F}}{\partial \tau} = -A_{ds}n\tau^{n-1} \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) - A_{df} \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right)$$

Once the value of τ is found, the strain rate values of Eqs. (5.86) and (5.87) can be computed and so can the respective effective viscosities:

$$\eta_{ds} = \frac{1}{2} A_{ds}^{1/n} \dot{\varepsilon}_{ds}^{\frac{1}{n}-1} \exp\left(\frac{Q_{ds} + pV_{ds}}{nRT}\right) \tag{5.88}$$

$$\eta_{df} = \frac{1}{2} A_{df}^{1/n} \exp\left(\frac{Q_{df} + pV_{df}}{RT}\right) \tag{5.89}$$

Their average effective viscosity $\tilde{\eta}_{eff}$ is given by

$$\tilde{\eta}_{eff} = \left(\frac{1}{\eta_{ds}} + \frac{1}{\eta_{df}}\right)^{-1}$$

Rather importantly, as we will see hereafter, the following variant is implemented in some codes (e.g. DOUAR , FANTOM , SOPALE , and probably many others) so as to bypass these costly Newton iterations:

1. compute η_{ds} and η_{df} with the *same* strainrate $\dot{\varepsilon}_T$, pressure and temperature values

2. average them by means of an harmonic average

In this case, we have

$$\dot{\varepsilon}_T = A_{df}\tau_{df} \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) \qquad \dot{\varepsilon}_T = A_{ds}\tau_{ds}^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right)$$

or,

$$\eta_{ds} = \frac{1}{2} A_{ds}^{1/n} \dot{\varepsilon}_T^{\frac{1}{n}-1} \exp\left(\frac{Q_{ds} + pV_{ds}}{nRT}\right) \tag{5.90}$$

$$\eta_{df} = \frac{1}{2} A_{df}^{1/n} \exp\left(\frac{Q_{df} + pV_{df}}{RT}\right) \tag{5.91}$$

We see that this simplification has consequences on the dislocation creep viscosity only.

A concrete example Let us consider a vertical section of upper mantle, from 660km depth to 30km depth. The lithosphere is assumed to be 90km thick. The temperature at the moho (the top of the domain) is set to 550C, 1330C at the LMB and 1380C at the bottom. A constant strainrate $\dot{\epsilon}_T = 10^{-15}\text{s}^{-1}$ is assumed. We assume that the pressure is lithostatic (for simplicity the density is taken to be constant at 3300kg/m^3). The temperature and pressure fields are shown hereunder:
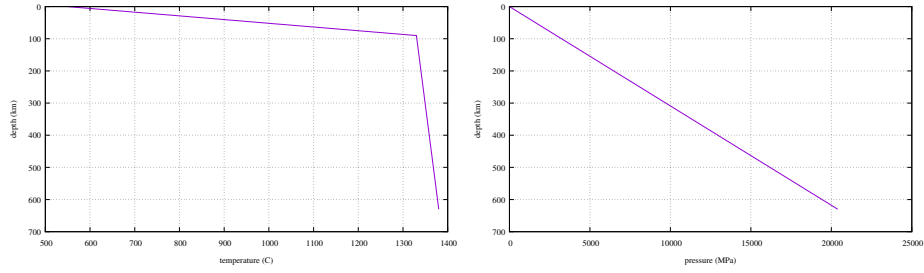
Material properties are taken from Karato & Wu (1993) [714]. The (fortran) code is available in `images/rheology/effvisc/`.

In what follows, the values obtained with Newton iterations are coined 'NR' and those obtained without are coined 'CHEAP'. The diffusion and dislocation creep viscosities can be computed for both algorithms and are shown hereunder (As mentioned earlier the diffusion creep viscosity is independent of strain rate so is the same for both):
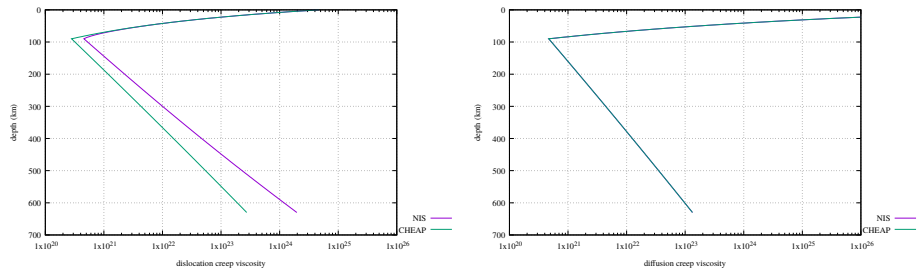
We can also plot the resulting effective viscosity η_{eff} for both approaches and we see that the differences are larger than 20%. This is shown here under on the left, alongside with the partitioning of the strain rate as a function of depth:
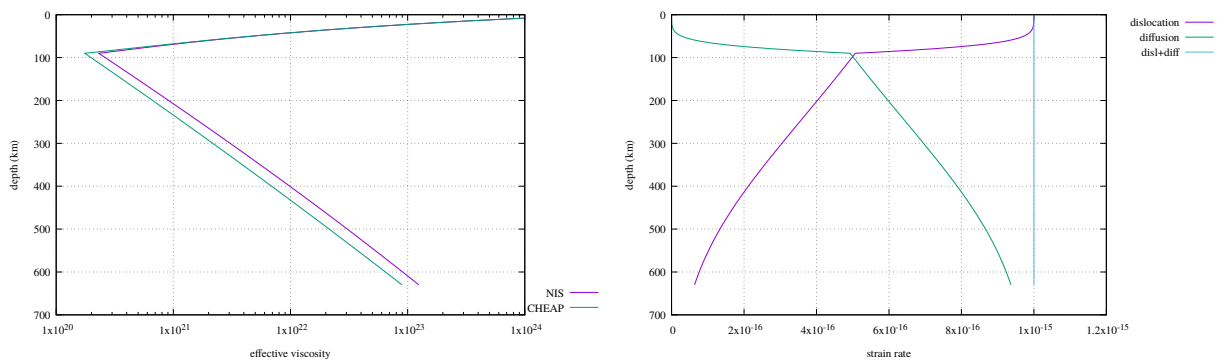
5.23.6 multiple viscous dampers and a plastic element
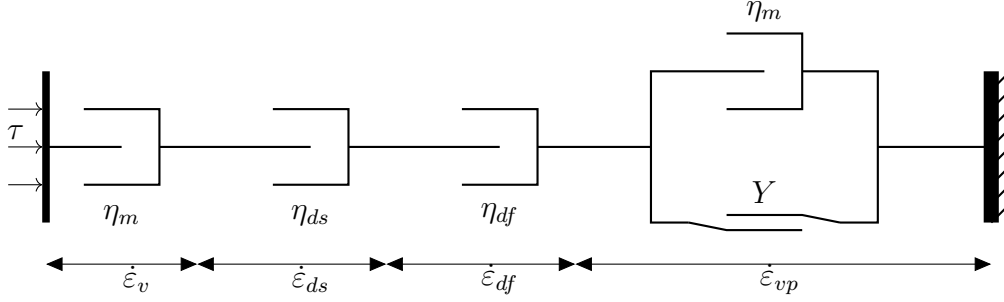
The algorithm goes then as follows:

1. Assume we know $\dot{\varepsilon}_T$ (from previous iteration), as well as the plasticity parameters Y (a constant in the case of von Mises, or a pressure-dependent quantity otherwise) and η_m.

2. We start by assuming that the plasticity 'block' is not active ($\dot{\varepsilon}_{vp} = 0$): we have then three dampers in series. We need their associated strain rates $\dot{\varepsilon}_{df}$ and $\dot{\varepsilon}_{ds}$ which are such that

$$\dot{\varepsilon}_T = \dot{\varepsilon}_v + \dot{\varepsilon}_{ds} + \dot{\varepsilon}_{df}$$

with

$$\dot{\varepsilon}_v = \frac{\tau}{2\eta_v} \tag{5.92}$$

$$\dot{\varepsilon}_{ds} = A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) \tag{5.93}$$

$$\dot{\varepsilon}_{df} = A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) \tag{5.94}$$

such that we are in fact looking for the stress value τ so that

$$\dot{\varepsilon}_T = A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) + A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) + \frac{\tau}{2\eta_v}$$

or, we must find the zero of the function $\mathcal{F}$:

$$\mathcal{F}(\tau) = \dot{\varepsilon}_T - A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) - A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) - \frac{\tau}{2\eta_v}$$

This equation can be solved with a Newton-Raphson algorithm and the iterations will be of the form:

$$\tau_{n+1} = \tau_n - \frac{\mathcal{F}(\tau_n)}{\mathcal{F}'(\tau_n)}$$

where the derivative of the function $\mathcal{F}$ with respect to τ reads:

$$\mathcal{F}'(\tau) = \frac{\partial \mathcal{F}}{\partial \tau} = -A_{df} \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right) - A_{ds}n\tau^{n-1} \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) - \frac{1}{2\eta_v}$$

Once the value of τ is found, the strain rate values of Eqs. (5.93), (5.94) and (5.92) can be computed and so can the respective effective viscosities:

$$\eta_{ds} = \frac{1}{2}A_{ds}^{1/n}\dot{\varepsilon}_{ds}^{\frac{1}{n}-1} \exp\left(\frac{Q_{ds} + pV_{ds}}{nRT}\right) \tag{5.95}$$

$$\eta_{df} = \frac{1}{2}A_{df}^{1/n} \exp\left(\frac{Q_{df} + pV_{df}}{RT}\right) \tag{5.96}$$

Their average effective viscosity $\tilde{\eta}_{eff}$ is given by

$$\tilde{\eta}_{eff} = \left(\frac{1}{\eta_{ds}} + \frac{1}{\eta_{df}} + \frac{1}{\eta_v}\right)^{-1}$$

3. if $\tau = 2\tilde{\eta}_{eff}\dot{\varepsilon}_T < Y$ the stress is below the yield stress value and the plasticity element is indeed not active. Use $\tilde{\eta}_{eff}$ in the material model.

4. if $\tau = 2\tilde{\eta}_{eff}\dot{\varepsilon}_T > Y$ the stress is above the yield value, which is not allowed. In this case the plastic element must be present and active and the viscous dampers are then in series with the (visco)plastic element. The formers deform with a strain rate $\dot{\varepsilon}_v$, $\dot{\epsilon}_{ds}$ and $\dot{\epsilon}_{df}$ while the latter with $\dot{\epsilon}_{vp}$ (all under the same tress τ) and we have $\dot{\varepsilon}_T = \dot{\varepsilon}_v + \dot{\varepsilon}_{ds} + \dot{\varepsilon}_{df} + \dot{\varepsilon}_{vp}$ so:

$$\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) - \dot{\varepsilon}_{df}(\tau) = \dot{\varepsilon}_{vp}$$

$$= \frac{\tau}{2\left(\frac{Y}{2\dot{\varepsilon}_{vp}} + \eta_m\right)}$$

$$\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) - \dot{\varepsilon}_{df}(\tau) = \frac{\tau}{2\left(\frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) + \dot{\varepsilon}_{df}(\tau))} + \right.}$$

$$2[\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) - \dot{\varepsilon}_{df}(\tau)]\left(\frac{Y}{2(\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) + \dot{\varepsilon}_{df}(\tau))} + \eta_m\right) = \tau$$

$$Y + 2(\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) - \dot{\varepsilon}_{df}(\tau))\eta_m = \tau$$

As before, we must find the zero of the function $\mathcal{F}$:

$$\mathcal{F}(\tau) = Y + 2[\dot{\varepsilon}_T - \dot{\varepsilon}_v(\tau) - \dot{\varepsilon}_{ds}(\tau) - \dot{\varepsilon}_{df}(\tau)]\eta_m - \tau$$

$$= Y + 2\left[\dot{\varepsilon}_T - \frac{\tau}{2\eta_v} - A_{ds}\tau^n \exp\left(-\frac{Q_{ds} + pV_{ds}}{RT}\right) - A_{df}\tau \exp\left(-\frac{Q_{df} + pV_{df}}{RT}\right)\right]\eta_m - \tau$$

Because dislocation creep involves the n-th power of the stress we will here also need to find the zero by means of a Newton-Raphson algorithm.

We have:

$$\frac{\partial\mathcal{F}}{\partial\tau} = \left[-\frac{1}{\eta_v} - 2\frac{\partial\dot{\varepsilon}_{ds}(\tau)}{\partial\tau} - 2\frac{\partial\dot{\varepsilon}_{df}(\tau)}{\partial\tau}\right]\eta_m - 1 \qquad (5.97)$$

$$\mathcal{F}(\tau)/2\eta_m = \frac{Y}{2\eta_m} + \dot{\varepsilon}_T - \frac{\tau}{2\eta_v} - A_{ds}(p,T)\tau^n - A_{df}(p,T)\tau - \frac{\tau}{2\eta_m} \qquad (5.98)$$

$$= -A_{ds}(p,T)\tau^n - \left(A_{df}(p,T) + \frac{1}{2\eta_v} - \frac{1}{2\eta_m}\right)\tau + \left(\frac{Y}{2\eta_m} + \dot{\varepsilon}_T\right) = 0 \quad (5.99)$$

Note that when $\eta_m = 0$ we logically recover $\tau = Y$ as the stress cannot exceed the yield strength Y.

Although this approach is probably the most consistent in terms of physics, the presence of the Newton-Raphson iterations makes it very expensive since this procedure is to be repeated for every quadrature point or every particle.

Let us consider a concrete example: we set $Y = 20$MPa, $\eta_v = 10^{25}$pascal, $\eta_m = 10^{20}$pascal. The domain is one-dimensional of depth 660km. The density is assumed to be constant at 3300kg m^{-3}. Dislocation and diffusion creep parameters are taken from Karato & Wu (1993) [714]. The temperature is linear is 20°C at the surface, 550°C at 30km depth, 1330celsius at 90km depth and 1380°C at the bottom. Pressure is assumed to be lithostatic. The python program and the gnuplot script are in *images/rheology/example*.

In the code I consider two cases: 'old' and 'new'. The latter is described above. 'old' goes as follows: loop over total strain rate values. Compute dislocation and diffusion creep viscosities with it. Compute harmonic average of these with linear viscosity. Compute deviatoric stress value. use it in dislocation and diffusion formulae to arrive at respective strainrates.
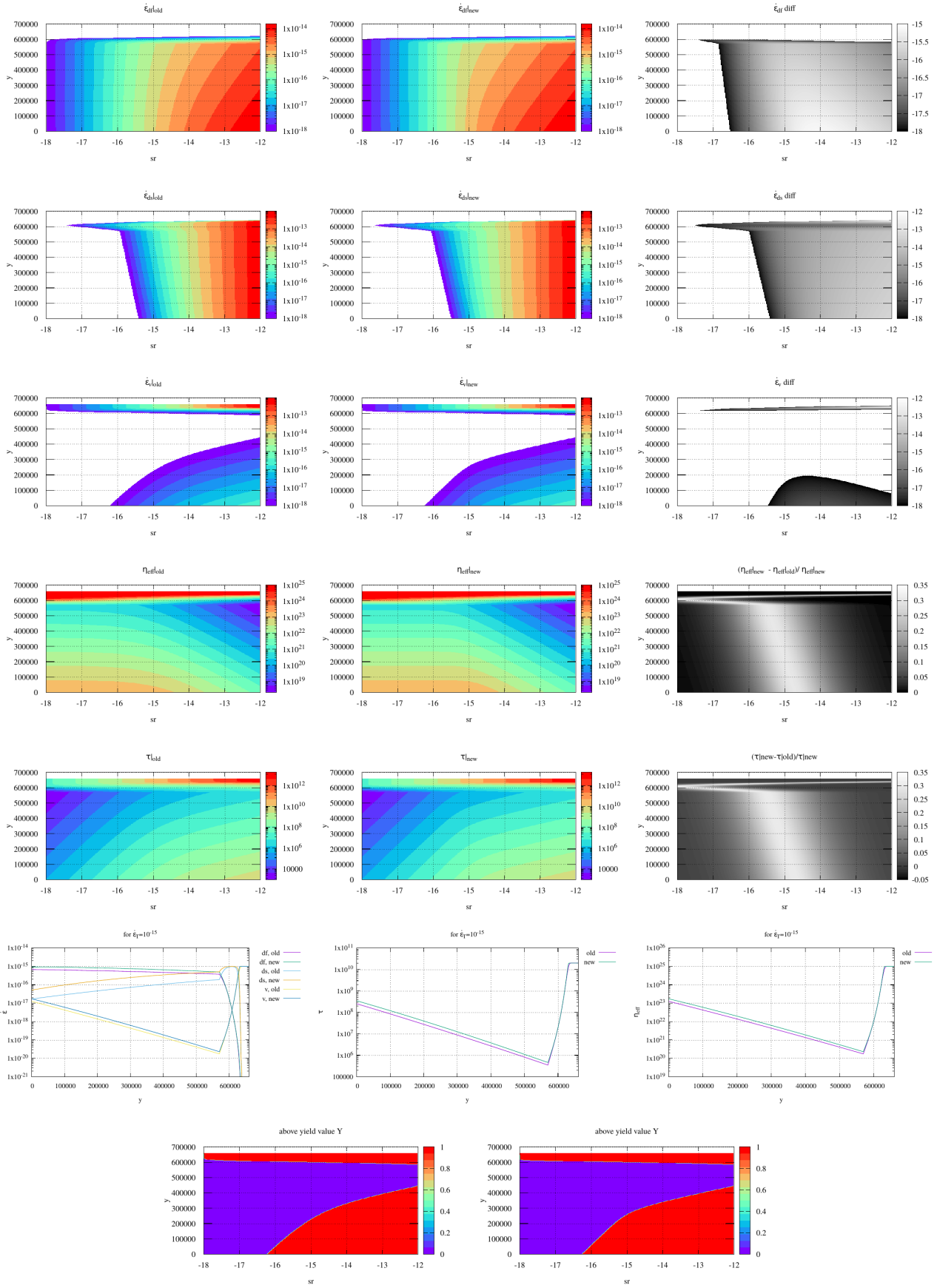
Viscous branch: (ds+df+v) 'old' stands for the old approach when $\dot{\varepsilon}_T$ was used for all mechanisms. 'new' stands for the new approach and the right strain rate decomposition.
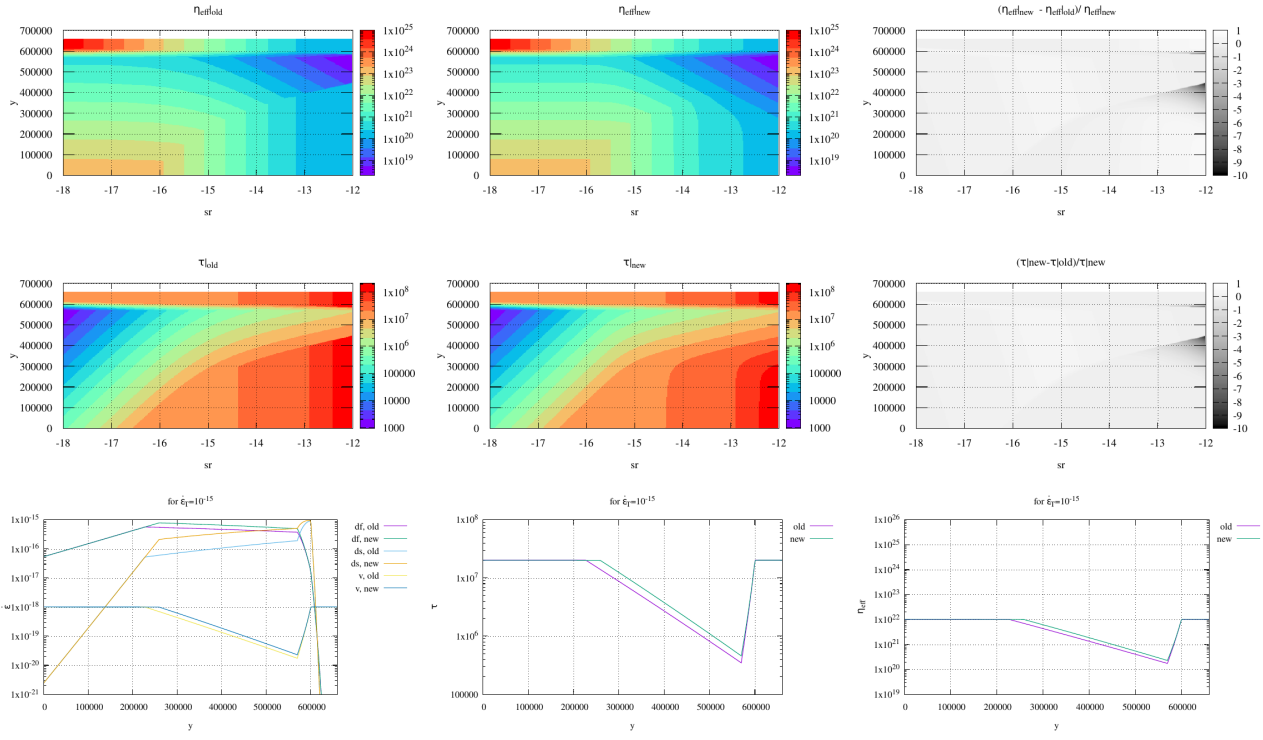
Visco-viscoplastic rheology: (ds+df+v+vp)

Remark. *Chenin et al. (2019) [247], base their rheological model on the additive decomposition of the following deviatoric strain rate tensor $\boldsymbol{\varepsilon}^d$:*

$$\boldsymbol{\varepsilon}^d = \boldsymbol{\varepsilon}^{el} + \boldsymbol{\varepsilon}^{pl} + \boldsymbol{\varepsilon}^{ds} + \boldsymbol{\varepsilon}^{df} + \boldsymbol{\varepsilon}^{pe}$$

where the five strain rate terms correspond respectively to the elastic, plastic, and viscous creep (dislocation, diffusion, peierls) contributions. This implies that all these elements are in series and the associated viscosities are then averaged with an harmonic mean. Rather interestingly, it is then stated that "this strain rate equation is nonlinear and solved locally on cell centroids and vertices in order to define the current effective viscosity and stress [1076]."

📖 Relevant Literature: Hobbs and Ord [613], Loret and Prevost [854], Hobbs, Mühlhaus, and Ord [612], Schmalholz, Podladchikov, and Schmid [1193], Lof and van den Boogaard [851], Andrés-Martínez, Pérez-Gussinyé, Armitage, and Morgan [25], and Elesin, Gerya, Artemieva, and Thybo [390]

5.24 Anisotropic viscosity

Following the paper by Lev and Hager (2008) [822], the anisotropic viscosity enters the equation of momentum through a 'correction' term added to the isotropic part of the constitutive equation relating stress and strain rate [968]:

$$\sigma_{ij} = -p\delta_{ij} + 2\eta_N\dot{\varepsilon}_{ij} - 2(\eta_N - \eta_S)\Lambda_{ijkl}\dot{\varepsilon}_{kl}$$

where η_N is the normal viscosity and η_S is the shear viscosity. The fourth order tensor Λ reflects the orientation of the directors in space, denoted by $\vec{n}$:

$$\Lambda_{ijkl} = \frac{1}{2}(n_i n_k \delta_{lj} + n_j n_k \delta_{il} + n_i n_l \delta_{kj} n_j n_l \delta_{ik}) - 2n_i n_j n_k n_l$$

154

Following [951, 968], the 'directors' are advected through the model and are analogous to particles. The directors are vector-particles pointing normal to the easy-glide plane or layer, thus defining the directions associated with η_N and η_S. In each time step of the calculation, the directors are advected and rotated by the flow, and in return determine the viscosity structure for the next time step [966].
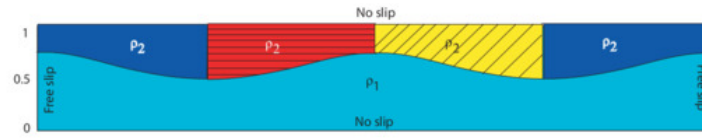
Figure 4. A schematic description of the model geometry and initial conditions. The colours denote the densities and rheologies: blue—isotropic, $\rho = 1$, $\eta_{iso} = 1$, red—anisotropic with horizontal fabric, $\rho = 1$, $\delta = 0.1$, yellow—anisotropic with dipping fabric, $\rho = 1$, $\delta = 10$, cyan—isotropic, $\rho = 0$, $\eta_{iso} = 1$). There is no slip on the top and bottom boundaries, and free slip is allowed along the side walls. The thickness of the top layer and the amplitude of the interface perturbation were exaggerated for clarity.
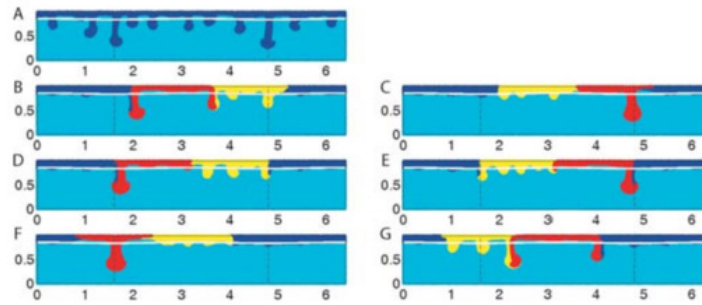
Figure 5. Material distribution in models with different configurations of initial anisotropic fabric taken after the fastest downwelling sinks over half the box depth. Panel A shows the results for an isotropic model. The black cosine curve at a depth of 0.15 marks the original interface between the dense and buoyant layers. The vertical dashed black lines show the deepest points of the original density interface, where the dense layer was thickest. Red material starts with a horizontal fabric; Yellow material starts with a fabric dipping at 45°. Blue materials are isotropic. Interestingly both panels (b) and (g), which start with distinctly different material arrangements, show large downwellings comprised of both anisotropic materials, while others do not.

Taken from Lev & Hager (2008) [822].

🧊 Relevant Literature:

- F.M. Richter and S.F. Daly. "Convection models having a multiplicy of large horizontal scales". In: *J. Geophys. Res.: Solid Earth* 83 (1978), pp. 4951–4956

- M. Saito and Y. Abe. "Consequences of anisotropic viscosity in the Earth's mantle (in Japanese, with English Abstract)". In: *Zisin* 37 (1984), pp. 237–245

- A. Vauchez, A. Tomassi, and G. Barroul. "Rheological heterogeneity, mechanical anisotropy and deformation of the continental lithosphere". In: *Tectonophysics* 296 (1998), pp. 61–86. DOI: 10.1016/S0040-1951(98)00137-1

- H.-B. Mühlhaus, L. Moresi, B. Hobbs, and F. Dufour. "Large amplitude folding in finely layered viscoelastic rock structures". In: *Pure Appl. Geophys.* 159 (2002), pp. 2311–2333

- H-B Mühlhaus, L Moresi, and M Čada. "Anisotropy model for mantle convection". In: *Computational Fluid and Solid Mechanics 2003*. 2003, pp. 1044–1046. DOI: 10.1016/B978-008044046-0.50255-4

- H-B Mühlhaus, L Moresi, and Miroslav Cada. "Emergent anisotropy and flow alignment in viscous rock". In: *Pure Appl. Geophys.* 161.11-12 (2004), pp. 2451–2463

- K. Michibayashi and D. Mainprice. "The role of pre-existing mechanical anisotropy on shear zone development within oceanic mantle lithosphere: an example from the Oman ophiolite". In: *J. Petrol.* 45(2) (2004), pp. 405–414

- L Moresi and H-B Mühlhaus. "Anisotropic viscous models of large-deformation Mohr–Coulomb failure". In: *Philosophical Magazine* 86.21-22 (2006), pp. 3287–3305. DOI: 10.1080/14786430500255419

- Hans Mühlhaus, Louis Moresi, Lutz Gross, and Joseph Grotowski. "The influence of non-coaxiality on shear banding in viscous-plastic materials". In: *Granular Matter* 12.3 (2010), pp. 229–238. DOI: 10.1007/s10035-010-0176-9

- Hans B Mühlhaus, Jingyu Shi, Louise Olsen-Kettle, and Louis Moresi. "Effects of a non-coaxial flow rule on shear bands in viscous-plastic materials". In: *Granular Matter* 13.3 (2011), pp. 205–210

- W. Sharples, L.N. Moresi, M. Velic, M.A. Jadamec, and D.A. May. "Simulating faults and plate boundaries with a transversely isotropic plasticity model". In: *Phys. Earth. Planet. Inter.* 252 (2016), pp. 77–90. DOI: 10.1016/j.pepi.2015.11.007

- J. Perry-Houts and L. Karlstrom. "Anisotropic viscosity and time-evolving lithospheric instabilities due to aligned igneous intrusions". In: *Geophy. J. Int.* 216.2 (2018), pp. 794–802. DOI: 10.1093/gji/ggy466

- Á Király, Clinton P Conrad, and LN Hansen. "Evolving viscous anisotropy in the upper mantle and its geodynamic implications". In: *Geochem. Geophys. Geosyst.* 21 (2020), e2020GC009159. DOI: 10.1029/2020GC009159

5.25 Rheology of the lithosphere
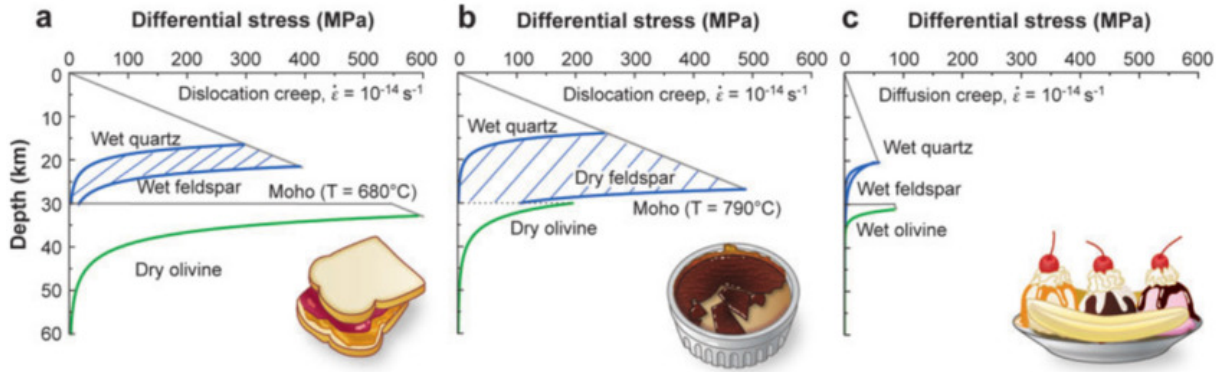
Schematic view of the three most common first order rheological models of the continental lithosphere under a strain rate of 10^{-14}s^{-1} . In all three models the upper crust has its frictional strength increased with pressure and depth. (a) The jelly sandwich model has a weak mid-lower crust and a strong mantle composed of dry olivine. (b) The crème brûlée model assumes that the mantle is weak, due to the presence of water and high temperature deformation, and the dry and brittle crust determines the strength of the lithosphere. (c) The banana split model assumes that the lithosphere as a whole has its strength greatly reduced due to various strain weakening and feedback processes [193]
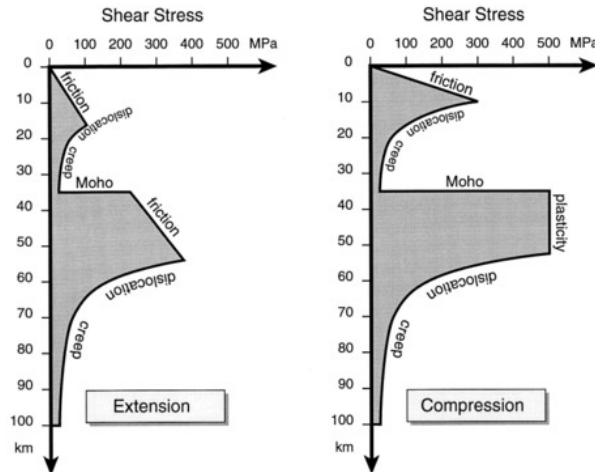
Taken from [99]. Typical vertical distribution of maximum shear stress in continental lithosphere undergoing compressional (right) or extensional (left) strain at 10^{-15}s. Friction controls level of shear stress in upper part of crust and sometimes in mantle lithosphere; then, below brittle/ductile transition, shear stress is controlled by thermally-activated dislocation creep.

Molnar [940] discusses the validity of the Brace-Goetze strength profiles. In particular, he has this to say about the power law parameters: *The uncertainty alone in Q alone renders calculated strengths uncertain by 10 times at temperatures of about 700C. Correspondingly, that uncertainty in Q is approximately equivalent to an uncertainty of about 100C in temperature.*

Wet Quartzite	upper crust [666, 1442]
	upper continental crust [737, 314]
	lower crust [666, 1442]
	ocean sediment [737]
Dry Olivine	lithosphere [647]
	sublithospheric mantle [647]
Dry Maryland Diabase	lower crust [1442, 1443]
	lower continental crust [737, 314]
	oceanic crust [1442, 737, 1443, 314]
Wet Olivine	continental mantle lithosphere [1442, 1443]
	oceanic mantle lithosphere [1442, 1443]
	sublithospheric mantle [1442, 737, 1443]
	mantle lithosphere [737]

📥 Relevant Literature[200, 193, 1110, 1109]

I need to talk about Byerlee's law. [211]

5.26 The Perzyna model

The Perzyna formulation is mentioned in `https://en.wikipedia.org/wiki/Viscoplasticity`. It is a rate-dependent plasticity formulation that was proposed in 1966 by Piotr Perzyna [1056]. See also `https://neml.readthedocs.io/en/dev/vp_flow/perzyna.html`.

In what follows I make use of the approach and notations of Zienkiewicz (1974) [1533] (and all the 1974-75 papers that follow) and the book by Owen & Hinton [1026].

The total strain (rate) is divided into two parts[12]:

$$\dot{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}}^e + \dot{\boldsymbol{\varepsilon}}^{vp}$$

where $\boldsymbol{\varepsilon}^e$ stands for the elastic strain tensor and $\boldsymbol{\varepsilon}^{vp}$ stands for the visco-plastic strain tensor.

The yield condition is given as

$$\mathcal{F}(\boldsymbol{\sigma}, \kappa) = \Psi(\boldsymbol{\sigma}, \dot{\boldsymbol{\varepsilon}}) - Y(\kappa) = 0$$

with $\mathcal{F} < 0$ denoting the purely elastic region, κ is a history-dependent hardening/softening parameter and $Y(\kappa)$ is a static yield stress. Ψ is a function of the stress and/or strain rate invariants.

We borrow from classical viscoplasticity theory (Perzyna (1966) [1056], Perzyna (1988) [1055]) the idea of a plastic potential defined as $\mathcal{Q}(\boldsymbol{\sigma})$ and write

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma \left\langle \phi\left(\mathcal{F}\right) \right\rangle \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} \tag{5.100}$$

where γ is a positive, possibly time-dependent fluidity parameter. Note that sometimes the pseudo-viscosity $\bar{\eta} = \gamma^{-1}$ is defined [1535] so that the equation above writes:

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \frac{1}{\bar{\eta}} \left\langle \phi\left(\mathcal{F}\right) \right\rangle \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} \tag{5.101}$$

$\mathcal{F}$ represents the plastic yield condition. $\phi(x)$ is a positive scalar-valued monotonic increasing function in the range $x > 0$ such that $\phi^{-1}(x)$ exists and possess similar properties in the same range. The notation $\langle \rangle$ denotes the Macaulay brackets[13] and stands for[14]

$$\langle \phi(x) \rangle = \phi(x) \quad \text{if} \quad x > 0$$
$$\langle \phi(x) \rangle = 0 \quad \text{if} \quad x \leq 0$$

If $\mathcal{Q} = \mathcal{F}$ then we speak of an *associative* law and if $\mathcal{Q} \neq \mathcal{F}$ we have a *non-associative* situation. The tensor $\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}}$ represents the direction of plastic flow and when $\mathcal{F} = \mathcal{Q}$ it is a vector directed normal to the yield surface at the stress point under consideration. This is potentially problematic in the case of the Tresca and Mohr-Coulomb yield surfaces since the normal is not well defined along the apices of the surfaces (see Section 7.6 of [1026]). In the non-associative case, the direction of plastic flow in the principal stress space during plastic flow is not the same as the direction of the vector normal to the yield surface.

In what follows we concentrate our attention on isotropic materials for which both $\mathcal{F}$ and $\mathcal{Q}$ can be defined in terms of stress invariants.

According to Zienkiewicz et al (1975) [1537]:

---

[12]Zienkiewicz [1533] adds a third term $\boldsymbol{\varepsilon}^0$ which stands for initial/autogenous strain such as due to temperature changes but I neglect it in what follows.

[13]`https://en.wikipedia.org/wiki/Macaulay_brackets`

[14]there is a difference between [1533](1974) and [1534](1974) wrt $>$ and $\geq$, and also a difference with wikipedia!

One of the main stumbling blocks of the classical plasticity theory lay in the universal assumption, based on Drucker's postulates (Drucker and Prager, 1952), that the plastic behaviour is 'associated'. With the use of Mohr-Coulomb type yield envelopes to define the limit between states of elasticity and of continuing irreversible deformation, the associated behaviour manifestly contradicted observation and gave excessive dilation. It became necessary therefore to extend plasticity ideas to a 'non-associated' form in which the plastic potential and yield surfaces are defined separately.

At the same time, it is worth remembering that these early studies mostly dealt with plasticity in metals, and later soils, but not kilometer-scale crustal layers.

Also, the Perzyna model is not the only one, see for instance the Duvaut-Lions viscoplastic model or the Consistency model [1438, 591].

We therefore need to look into the derivative of the plastic potential $\mathcal{Q}$ with respect to the stress tensor. Since the potential is expressed as a function of the stress invariants $\mathcal{I}_1(\boldsymbol{\sigma})$, $\mathcal{I}_2(\boldsymbol{\tau})$ and $\theta_L(\boldsymbol{\tau})$, we then have[15]:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} &= \frac{\partial}{\partial \boldsymbol{\sigma}} \mathcal{Q}\left(\mathcal{I}_1(\boldsymbol{\sigma}), \mathcal{I}_2(\boldsymbol{\tau}), \theta_{\mathrm{L}}(\boldsymbol{\tau})\right) \\
&= \frac{\partial \mathcal{Q}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + \frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \frac{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&= \frac{\partial \mathcal{Q}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + \frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&\quad - \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \left[-\frac{3}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{5/2}} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right] \\
&= \frac{\partial \mathcal{Q}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + \left(\frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} + \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \frac{3}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{5/2}} \right) \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&\quad - \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&= C_1 \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + C_2 \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + C_3 \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \qquad (5.102)
\end{aligned}
$$

i.e.

$$
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} = C_1 \boldsymbol{a}_1 + C_2 \boldsymbol{a}_2 + C_3 \boldsymbol{a}_3 = C_1 \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + C_2 \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + C_3 \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \qquad (5.103)
$$

---

[15]The derivative of the Lodé angle was obtained in Section 4.2.2

where the $C_{1,2,3}$ coefficients depend on the plastic potential $\mathcal{Q}$ and the stress invariants as follows:

$$C_1 = \frac{\partial \mathcal{Q}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \tag{5.104}$$

$$C_2 = \frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} + \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \frac{3}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{5/2}}$$

$$= \frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} - \frac{1}{2} \frac{\tan 3\theta_{\mathrm{L}}}{\mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})}$$

$$= \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \left(\frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} - \frac{\tan 3\theta_{\mathrm{L}}}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \right)$$

$$C_3 = -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \frac{\partial \mathcal{Q}}{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})} \tag{5.105}$$

These are identical to those of Eq. (7.71) in Owen & Hinton[16]:

$$\boxed{\begin{aligned} C_1 &= \frac{\partial F}{\partial J_1}, \quad C_2 = \left(\frac{\partial F}{\partial (J_2')^{1/2}} - \frac{\tan 3\theta}{(J_2')^{1/2}} \frac{\partial F}{\partial \theta} \right), \\ C_3 &= \frac{-\sqrt{3}}{2\cos 3\theta} \frac{1}{(J_2')^{3/2}} \frac{\partial F}{\partial \theta}. \end{aligned} \tag{7.71}}$$

Note that we already have established (see Section 4.7) that

$$\boldsymbol{a}_1 = \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \mathbf{1}$$

$$\boldsymbol{a}_2 = \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} = \boldsymbol{\tau}$$

$$\boldsymbol{a}_3 = \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} = \boldsymbol{\tau} \cdot \boldsymbol{\tau} - \frac{2}{3} \mathcal{I}_2(\boldsymbol{\tau}) \mathbf{1}$$

with (we will need these soon)

$$\mathrm{Tr}[\boldsymbol{a}_1] = \mathrm{tr}\left[\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} \right] = 3 \tag{5.106}$$

$$\mathrm{Tr}[\boldsymbol{a}_2] = \mathrm{tr}\left[\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right] = 0 \tag{5.107}$$

$$\mathrm{Tr}[\boldsymbol{a}_3] = \mathrm{tr}\left[\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right] = \mathrm{tr}[\boldsymbol{\tau} \cdot \boldsymbol{\tau}] - 2\mathcal{I}_2(\boldsymbol{\tau}) = 2\mathcal{I}_2(\boldsymbol{\tau}) - 2\mathcal{I}_2(\boldsymbol{\tau}) = 0 \tag{5.108}$$

The generic form of the plastic potential derivative then reads

$$\boxed{\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} = C_1 \mathbf{1} + C_2 \boldsymbol{\tau} + C_3 \left(\boldsymbol{\tau} \cdot \boldsymbol{\tau} - \frac{2}{3} \mathcal{I}_2(\boldsymbol{\tau}) \mathbf{1} \right)} \tag{5.109}$$

The momentum conservation equation that we solve is

$$-\vec{\nabla} p + \vec{\nabla} \cdot \left[2\eta \dot{\boldsymbol{\varepsilon}}^d(\vec{\mathrm{v}}) \right] + \rho \vec{g} = \vec{0}$$

---

[16]This is not exactly true: the factor $\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}$ is absent in their Eq. (7.71) but it is to be found in their Eq. (7.70).

so we need the *deviatoric* strain rate tensor. We here assume for simplicity that there is only a visco-plastic element in the system (or that the other mechanisms are deviatoric), i.e. $\dot{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}}^{vp}$. Then

$$
\begin{aligned}
\dot{\boldsymbol{\varepsilon}}^d &= \dot{\boldsymbol{\varepsilon}}^{vp} - \frac{1}{3}\mathrm{tr}[\dot{\boldsymbol{\varepsilon}}^{vp}]\mathbf{1} \\
&= \gamma\langle\phi(\mathcal{F})\rangle\left\{(C_1\boldsymbol{a}_1 + C_2\boldsymbol{a}_2 + C_3\boldsymbol{a}_3) - \frac{1}{3}\mathrm{tr}\,(C_1\boldsymbol{a}_1 + C_2\boldsymbol{a}_2 + C_3\boldsymbol{a}_3)\,\mathbf{1}\right\} \\
&= \gamma\langle\phi(\mathcal{F})\rangle\left\{(C_1\boldsymbol{a}_1 + C_2\boldsymbol{a}_2 + C_3\boldsymbol{a}_3) - \frac{1}{3}C_1 3\mathbf{1}\right\} \\
&= \gamma\langle\phi(\mathcal{F})\rangle\left\{\left(C_1\mathbf{1} + C_2\boldsymbol{\tau} + C_3(\boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1})\right) - C_1\mathbf{1}\right\} \\
&= \gamma\langle\phi(\mathcal{F})\rangle\left(C_2\boldsymbol{\tau} + C_3(\boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1})\right) \tag{5.110}
\end{aligned}
$$

The $C_{1,2,3}$ coefficients have been computed in Sections 5.11, 5.12, 5.14 and 5.13, and are summarized below:

	C_1	$C_2(\times\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}})$	C_3
Tresca	0	$2\cos\theta_L(1 + 2\tan\theta_L\tan 3\theta_L)$	$\frac{\sqrt{3}}{\mathcal{I}_2(\boldsymbol{\tau})}\frac{\sin\theta_L}{\cos 3\theta_L}$
von Mises	0	$\sqrt{3}$	0
Mohr-Coulomb	$\frac{1}{3}\sin\phi$	$\cos\theta_L\left[(1 + 2\tan\theta_L\tan 3\theta_L) + \frac{1}{\sqrt{3}}\sin\phi(2\tan 3\theta_L - \tan\theta_L)\right]$	$\frac{\sqrt{3}\sin\theta_L + \sin\phi\cos\theta_L}{2\mathcal{I}_2(\boldsymbol{\tau})\cos 3\theta_L}$
Drucker-Prager	α	1	0

The difference with the table below taken from Owen & Hinton [1026] is highlighted in blue.

Table 7.1	Constants defining the yield surface in a form suitable for numerical analysis.		
Yield Criterion	C_1	C_2	C_3
Tresca	0	$2\cos\theta(1 + \tan\theta\tan 3\theta)$	$\frac{\sqrt{3}}{J_2'}\frac{\sin\theta}{\cos 3\theta}$
Von Mises	0	$\sqrt{3}$	0
Mohr–Coulomb	$\frac{1}{3}\sin\phi$	$\cos\theta[(1 + \tan\theta\tan 3\theta) + \sin\phi(\tan 3\theta - \tan\theta)/\sqrt{3}]$	$\frac{(\sqrt{3}\sin\theta + \cos\theta\sin\phi)}{(2J_2'\cos 3\theta)}$
Drucker–Prager	α	1.0	0

Taken from [1026]. This table supposedly presents all three $C_{1,2,3}$ coefficients for all four plastic potentials/yield functions (associative plasticity). This is however not the case: the C_2 column is not C_2 but $\partial\mathcal{F}/\partial\sqrt{I_2}$!

Also, the continuity equation for incompressible flow contains the divergence of the velocity field,

and in this case

$$
\begin{aligned}
\vec{\nabla} \cdot \vec{v} &= \mathrm{tr}[\dot{\boldsymbol{\varepsilon}}^{vp}] \\
&= \mathrm{tr}\left[\gamma\left\langle\phi\left(\mathcal{F}\right)\right\rangle\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}}\right] \\
&= \gamma\left\langle\phi\left(\mathcal{F}\right)\right\rangle\mathrm{tr}\left[C_1\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + C_2\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + C_3\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}}\right] \\
&= \gamma\left\langle\phi\left(\mathcal{F}\right)\right\rangle\mathrm{tr}\left[C_1\mathbf{1} + C_2\boldsymbol{\tau} + C_3\left(\boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1}\right)\right] \\
&= \gamma\langle\phi(\mathcal{F})\rangle 3C_1
\end{aligned}
\tag{5.111}
$$

I find it difficult to wrap my head around this as the continuity equation is usually derived by other means. If C_1 is not zero, then dilation occurs, the material is not incompressible so density should also change... On the other hand this is a 'problem' only when the bracket is nonzero, i.e. when plasticity is activated. If the Macaulay bracket is zero, then we recover the zero divergence constraint.

5.26.1   von Mises plasticity following Zienkiewicz (1975)

What follows is borrowed from Zienkiewicz (1975) [1532].
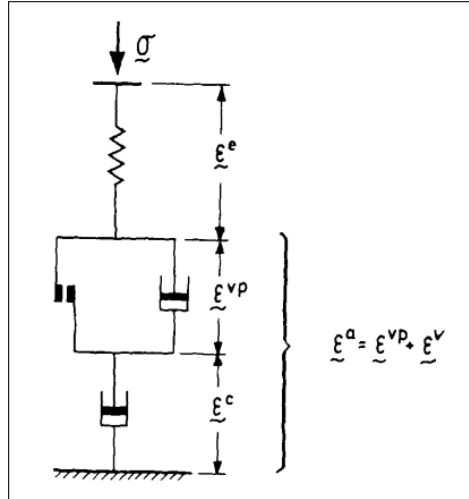
Taken from Zienkiewicz (1975) [1532].

We start from section 13.4.2 of the paper with the Perzyna formulation of the plastic strain[17].

$$
\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma\langle\phi(\mathcal{F})\rangle\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}}
$$

Associative plasticity is used, i.e. $\mathcal{F}^{\mathrm{vM}} = \mathcal{Q}^{\mathrm{vM}}$, and the von Mises yield criterion is $\mathcal{F}^{\mathrm{vM}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y$ so

$$
\begin{aligned}
\dot{\boldsymbol{\varepsilon}}^{vp} &= \gamma\left\langle\phi\left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y\right)\right\rangle\frac{\partial\left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y\right)}{\partial \boldsymbol{\sigma}} \\
&= \gamma\left\langle\phi\left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y\right)\right\rangle\frac{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}{\partial \boldsymbol{\sigma}} \\
&= \gamma\left\langle\phi\left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y\right)\right\rangle\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\frac{\partial\mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}}
\end{aligned}
\tag{5.112}
$$

---

[17]from which I have removed the unnecessary/uncommon $\sqrt{3}$ terms

Using results of Section 4.7 for the partial derivative of the second invariant we find[18]:

$$\dot{\boldsymbol{\varepsilon}}^{vp} \;=\; \gamma \Big\langle \phi \Big(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y \Big) \Big\rangle \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau} \tag{5.113}$$

which we can also write[19]

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \frac{1}{2\eta} \boldsymbol{\tau} \qquad \text{with} \qquad \frac{1}{2\eta} = \gamma \Big\langle \phi(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y) \Big\rangle \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}$$

Note that it here follows that the flow is incompressible since the visco-plastic strain rate tensor is proportional to the deviatoric stress tensor so it is deviatoric itself!

From the definition of the second moment invariant:

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2}(2\eta\dot{\boldsymbol{\varepsilon}}^{vp}) : (2\eta\dot{\boldsymbol{\varepsilon}}^{vp}) = 4\eta^2 \left(\frac{1}{2}\dot{\boldsymbol{\varepsilon}}^{vp} : \dot{\boldsymbol{\varepsilon}}^{vp} \right) = 4\eta^2 \mathcal{I}_2 \dot{\varepsilon}_e^{vp}$$

from which η can be found as a function of strain rates and hence $\boldsymbol{\Gamma}(\dot{\boldsymbol{\varepsilon}})$ becomes available (WHAT IS GAMMA?!). Note that annoyingly the author defines the second invariant as $2\dot{\boldsymbol{\varepsilon}} : \dot{\boldsymbol{\varepsilon}}$ in Eq. (13.50) of the paper.

It follows that

$$\tau_e = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} = 2\eta\dot{\varepsilon}_e^{vp}$$

Then we drop the $\langle \cdot \rangle$ as we assume to be above yield and we also assume a power-law form $\phi(\mathcal{F}) = \mathcal{F}^n$ so that we can solve explicitly for η:

$$\begin{aligned}
\frac{1}{2\eta} &= \gamma \Big\langle \phi(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y) \Big\rangle \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \\
\frac{1}{2\eta} &= \gamma \left(2\eta\sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}})} - Y \right)^n \frac{1}{2\, 2\eta\sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}})}} \\
\frac{1}{2\eta} &= \gamma \left(2\eta\dot{\varepsilon}_e - Y \right)^n \frac{1}{2\, 2\eta\dot{\varepsilon}_e} \\
2\dot{\varepsilon}_e &= \gamma(2\eta\dot{\varepsilon}_e - Y)^n \\
2\dot{\varepsilon}_e/\gamma &= (2\eta\dot{\varepsilon}_e - Y)^n \\
(2\dot{\varepsilon}_e/\gamma)^{1/n} &= 2\eta\dot{\varepsilon}_e - Y \\
\eta &= \frac{Y + (2\dot{\varepsilon}_e/\gamma)^{1/n}}{2\dot{\varepsilon}_e} \tag{5.114}
\end{aligned}$$

This form is convenient for plastic, visco-plastic and creep phenomena[20]. This can be re-written

$$\begin{aligned}
\eta &= \frac{Y + (2\dot{\varepsilon}_e/\gamma)^{1/n}}{2\dot{\varepsilon}_e} \\
&= \frac{Y}{2\dot{\varepsilon}_e} + \frac{(2\dot{\varepsilon}_e/\gamma)^{1/n}}{2\dot{\varepsilon}_e} \\
&= \frac{Y}{2\dot{\varepsilon}_e} + \frac{(2/\gamma)^{1/n}}{2} \dot{\varepsilon}_e^{\frac{1}{n}-1} \\
&= \frac{Y}{2\dot{\varepsilon}_e} + \frac{1}{2}(\gamma/2)^{-1/n} \dot{\varepsilon}_e^{\frac{1}{n}-1} \tag{5.115}
\end{aligned}$$

---

[18]This is the same equation as Eq. (14) of Zienkiewicz et al (1978) [1538]

[19]There most likely is a confusion in the paper between σ and τ there.

[20]Down to various $\sqrt{2}$ or $\sqrt{3}$ coefficients here or there, it is also to be found in Vilotte etal [1421, 1422, 1423]

We often use dislocation creep/power-law rheologies and these yield an effective viscosity $\frac{1}{2}A^{-1/n}\dot{\varepsilon}_e^{\frac{1}{n}-1}$ (the temperature and/or pressure-dependent exponential has been omitted for simplicity - it is a power law rheology). The equation above is then the sum of the 'plastic viscosity' and the 'viscous creep viscosity' – which corresponds to a dashpot and a plastic element in parallel. Note that the expression above is very similar to the one for Bingham or Herschel-Bulkley visco-plastic models.

If $n = 1$ then we find (as in Vilotte et al (1982) [1421] (1982))

$$\eta \;\; = \;\; \frac{Y}{2\dot{\varepsilon}_e} + \frac{1}{\gamma} \tag{5.116}$$

and γ is then the inverse of the (linear) viscosity of the dashpot. Also if $n = 1$ then $\bar{\eta} = \gamma^{-1}$.

For pure plasticity then $\gamma \to \infty$ and we have here simply

$$\eta = \frac{Y}{2\dot{\varepsilon}_e}$$

As stated in Vilotte etal (1982) [1421]:

> The plastic flow law of Eq. (5.115) permits us to represent in a single expression both the rigid-perfectly plastic flow ($\gamma \to \infty$) and the common power creep law without plastic limit ($Y = 0$) (usually referred to as the Norton-Hoff law)[21].

Vilotte [1421] then explain that the fluidity γ can depend on temperature T in the form

$$\gamma = \gamma_0 \exp(-Q/RT)$$

In this case Eq. (5.115) becomes

$$
\begin{aligned}
\eta \;\; &= \;\; \frac{Y}{2\dot{\varepsilon}_e} + \frac{1}{2}(\gamma_0 \exp(-Q/RT)/2)^{-1/n}\dot{\varepsilon}_e^{\frac{1}{n}-1} \\
&= \;\; \frac{Y}{2\dot{\varepsilon}_e} + \frac{1}{2}(\gamma_0/2)^{-1/n}\dot{\varepsilon}_e^{\frac{1}{n}-1}\exp(Q/nRT) \\
&= \;\; \frac{Y}{2\dot{\varepsilon}_e} + \frac{1}{2}A^{-1/n}\dot{\varepsilon}_e^{\frac{1}{n}-1}\exp(Q/nRT)
\end{aligned} \tag{5.117}
$$

where we have defined $A = \gamma_0/2$, which is the sum of the 'plastic viscosity' and a 'dislocation creep viscosity'.

In conclusion we find that this formulation allows us to represent linear, power law, perfectly plastic and visco-plastic materials. Also, we know that the additional term $\bar{\eta} = \gamma^{-1}$ introduces a length scale in the shear bands by limiting the viscosity value in said shear bands.

Remark. *If we use a non-associative plasticity then often $\mathcal{Q} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$. and the formulation of the previous section remains valid. In that case we have $C_1 = 0$ and $C_3 = 0$ which allows to easily arrive at a relationship of the type $\dot{\boldsymbol{\varepsilon}}^{vp} = \frac{1}{2\eta}\boldsymbol{\tau}$ where η is a scalar viscosity.*

However, if $\mathcal{Q}$ is such that $C_3 \neq 0$, then we have a problem because even by setting $n = 1$ I do not know how to arrive at a scalar viscosity, and even thinking of η as a tensor then I am stuck, see Section about Choi & Petersen (2015) hereafter.

---

[21]https://en.wikipedia.org/wiki/Viscoplasticity

5.26.2 Dissecting Choi & Petersen (2015)

In the paper the author state:

> In this paper, we confirm that using an associated flow rule is a straightforward way of acquiring shear band orientations tightly bound around the Coulomb angle in conventional numerical tectonic models. We further provide a simple analysis of why an associated flow rule constrains shear band orientations more tightly than a non- associated counterpart.

The original paper [255] is in 2D and focuses on the Mohr-Coulomb criterion. The authors state that the mass conservation equation should be

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = R = 2 \sin \psi \, \dot{\varepsilon}^p$$

where where R is the dilation rate, Ψ is the dilation angle and $\dot{\varepsilon}^p$ is the square root of the second invariant of the deviatoric plastic strain rate tensor.

After multiple reads of their paper, I originally had many questions:

- where does this dilation rate R come from ?

- after reading *many* papers or textbooks on plasticity I cannot see a factor 2 in an equation anymore without re-deriving it from scratch with a coherent set of notations (preferably mine in fieldstone).

- is this relationship still valid in 3D?

- is it the same term for Drucker-Prager ?

Time to investigate ...

Let us first look at their Eq. (3) in which the MC yield function is given by the function f:

$$f = \sigma_1 - N_\phi \sigma_3 - 2\sqrt{N_\phi} c$$

where σ_1 and σ_3 are the greatest and the least principal stress, $N_\phi = (1 + \sin \phi)/(1 - \sin \phi)$. This is a somewhat unusual formulation in the geodynamics community.

Let us then start with the MC yield criterion[22]

$$\tau_m = \sigma_m \sin \phi + c \cos \phi \tag{5.118}$$

which means that compression is assumed to be positive (the opposite as in fieldstone) and where τ_m is the magnitude of the shear stress, σ_m is the normal stress, c is the intercept of the failure envelope with the τ axis, and ϕ is the slope of the failure envelope. The quantity c is called the cohesion and the angle ϕ is called the angle of internal friction. We have

$$\sigma_m = \frac{1}{2}(\sigma_1 + \sigma_3)$$

and

$$\tau_m = \frac{1}{2}(\sigma_1 - \sigma_3)$$

---

[22]https://en.wikipedia.org/wiki/Mohr-Coulomb_theory

Inserting these into Eq. (5.118)

$$\frac{1}{2}(\sigma_1 - \sigma_3) = \frac{1}{2}(\sigma_1 + \sigma_3)\sin\phi + c\cos\phi \tag{5.119}$$

which can be reworked as follows:

$$\sigma_1 - \frac{1 + \sin\phi}{1 - \sin\phi}\sigma_3 - 2\frac{\cos\phi}{1 - \sin\phi}c = 0$$

The third term can further be modified as follows:

$$\frac{\cos\phi}{1 - \sin\phi} = \frac{\sqrt{1 - \sin^2\phi}}{\sqrt{(1 - \sin\phi)^2}} = \frac{\sqrt{(1 - \sin\phi)(1 + \sin\phi)}}{\sqrt{(1 - \sin\phi)^2}} = \sqrt{\frac{1 + \sin\phi}{1 - \sin\phi}}$$

Finally, we define N_ϕ as follows

$$N_\phi = \frac{1 + \sin\phi}{1 - \sin\phi}$$

so that the yield condition becomes:

$$\sigma_1 - N_\phi\sigma_3 - 2\sqrt{N_\phi}\,c = 0$$

which is Eq. 3 of the article by Choi & Petersen [255].

They also define the plastic potential as

$$g = \sigma_1 - N_\psi\sigma_3 = \tau_m - \sigma_m\sin\psi$$

We start again from the M-C criterion (in this case σ_2 replaces σ_3):

$$\frac{1}{2}(\sigma_1 - \sigma_2) = -\frac{1}{2}(\sigma_1 + \sigma_2)\sin\phi + c\cos\phi \tag{5.120}$$

In the case of incompressible flow I have established in Section 4.5 that

$$\frac{\sigma_1 + \sigma_2}{2} = \frac{\sigma_{xx} + \sigma_{yy}}{2} = \frac{1}{2}\mathcal{I}_1(\boldsymbol{\sigma}) \tag{5.121}$$

$$\frac{\sigma_1 - \sigma_2}{2} = \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{5.122}$$

so that we now have

$$\mathcal{F}^{\text{MC}} = \frac{1}{2}\mathcal{I}_1(\boldsymbol{\sigma})\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - c\cos\phi$$

and then the plastic potential $\mathcal{Q}$ is given by

$$\mathcal{Q}^{\text{MC}} = \frac{1}{2}\mathcal{I}_1(\boldsymbol{\sigma})\sin\psi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$$

We will need $\partial\mathcal{Q}/\partial\boldsymbol{\sigma}$. By applying the chain rule we can write

$$\begin{aligned}
\frac{\partial\mathcal{Q}}{\partial\boldsymbol{\sigma}} &= \frac{\partial\mathcal{Q}}{\partial\mathcal{I}_1(\boldsymbol{\sigma})}\frac{\partial\mathcal{I}_1(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}} + \frac{\partial\mathcal{Q}}{\partial\mathcal{I}_2(\boldsymbol{\tau})}\frac{\partial\mathcal{I}_2(\boldsymbol{\tau})}{\partial\boldsymbol{\sigma}} \\
&= \frac{\partial\mathcal{Q}}{\partial\mathcal{I}_1(\boldsymbol{\sigma})}\frac{\partial\mathcal{I}_1(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}} + \frac{\partial\mathcal{Q}}{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\frac{\partial\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}{\partial\mathcal{I}_2(\boldsymbol{\tau})}\frac{\partial\mathcal{I}_2(\boldsymbol{\tau})}{\partial\boldsymbol{\sigma}} \\
&= \frac{1}{2}\sin\psi\,\mathbf{1} + \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\boldsymbol{\tau}
\end{aligned} \tag{5.123}$$

Ultimately we would like to be able to write $\dot{\boldsymbol{\varepsilon}}^{vp} = \boldsymbol{\tau}/(2\eta)$ where η is the 'viscoplastic' viscosity. However, as opposed to Zienkiewicz (1975) in the previous section, the term $\partial \mathcal{Q}/\partial \boldsymbol{\sigma}$ is not directly/only proportional to the deviatoric stress $\boldsymbol{\tau}$ and we have instead:

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma \langle \phi(\mathcal{F}^{\mathrm{MC}}) \rangle \left(\frac{1}{2} \sin \psi \, \mathbf{1} + \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau} \right) \tag{5.124}$$

Right away we note that the strain rate tensor above is not deviatoric, i.e. the flow is not incompressible. Rather conveniently, the M-C criterion in plane strain can also be cast $\mathcal{F}^{\mathrm{MC}} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y = \tau_e - Y$ as in the von Mises case, albeit with $Y = -\frac{1}{2}\mathcal{I}_1(\boldsymbol{\sigma})\sin \phi + c\cos \phi$. Assuming $\phi(x) = x$ for convenience here and the argument of the brackets is positive,

$$\begin{aligned}
\dot{\boldsymbol{\varepsilon}}^{vp} &= \gamma \left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y \right) \left(\frac{1}{2} \sin \psi \, \mathbf{1} + \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau} \right) \\
&= \gamma \left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y \right) \frac{1}{2} \sin \psi \, \mathbf{1} + \gamma \left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y \right) \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau} \\
&= \gamma (\tau_e - Y) \frac{1}{2} \sin \psi \, \mathbf{1} + \gamma (\tau_e - Y) \frac{1}{2\tau_e} \boldsymbol{\tau}
\end{aligned} \tag{5.125}$$

If we follow the procedure of Zienkiewicz (1975) of the previous section, then the deviatoric part of the equation above would yield a viscosity

$$\eta = \frac{Y}{2\dot{\varepsilon}_e^{vp}} + \frac{1}{\gamma} \quad \Rightarrow \quad \tau_e = 2\eta\dot{\varepsilon}_e^{vp} = Y + \gamma^{-1}2\dot{\varepsilon}_e^{vp} \quad \Rightarrow \quad \tau_e - Y = \gamma^{-1}2\dot{\varepsilon}_e^{vp} \tag{5.126}$$

If we insert this in Eq. (5.125):

$$\begin{aligned}
\dot{\boldsymbol{\varepsilon}}^{vp} &= \gamma\gamma^{-1}2\dot{\varepsilon}_e^{vp} \frac{1}{2} \sin \psi \, \mathbf{1} + \frac{1}{2\eta} \boldsymbol{\tau} \\
&= \dot{\varepsilon}_e^{vp} \sin \psi \, \mathbf{1} + \frac{1}{2\eta} \boldsymbol{\tau}
\end{aligned} \tag{5.127}$$

Assuming that the total strain rate is the sum of the strain rates associated to the various deformation mechanisms, and that all other deformation mechanisms are deviatoric, then

$$div(\vec{\mathbf{v}}) = \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} = 2\dot{\varepsilon}_e^{vp} \sin \psi$$

This is identical to the dilation rate of Choi and Petersen [255]!

For implementation details in a FEM code, please check STONE 39.

5.26.3   my take on this in 3D for Drucker-Prager

I have established in Section 5.14 that in the general 3D case

$$\mathcal{F}^{\mathrm{DP}} = \alpha(\phi, c)\mathcal{I}_1(\boldsymbol{\sigma}) + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} + k(\phi, c) = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y \tag{5.128}$$

with α and k being functions of the cohesion c and angle of friction ϕ (but not from the stress). Then the plastic potential is

$$\mathcal{Q}^{\mathrm{DP}} = \alpha(\psi, c)\mathcal{I}_1(\boldsymbol{\sigma}) + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})} \tag{5.129}$$

where ψ is the dilation angle. We then have

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} &= \frac{\partial \mathcal{Q}}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + \frac{\partial \mathcal{Q}}{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{\partial \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&= \alpha(\psi, c)\, \mathbf{1} + \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau}
\end{aligned}
\tag{5.130}
$$

Then

$$
\begin{aligned}
\dot{\boldsymbol{\varepsilon}}^{vp} &= \gamma\left(\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y\right)\left(\alpha(\psi, c)\, \mathbf{1} + \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau}\right) \\
&= \gamma\left(\tau_e - Y\right)\alpha(\psi, c)\, \mathbf{1} + \gamma\left(\tau_e - Y\right)\frac{1}{2\tau_e} \boldsymbol{\tau}
\end{aligned}
\tag{5.131}
$$

Using again $\tau_e - Y = \gamma^{-1} 2\dot{\varepsilon}_e^{vp}$ as in the 2D case we arrive finally

$$
div(\vec{v}) = \mathrm{tr}[\dot{\boldsymbol{\varepsilon}}^{vp}] = \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} = 6\alpha(\psi, c)\dot{\varepsilon}_e^{vp}
$$

Since k does not depend on stress, the only difference between the associative and the non-associative case is whether $\phi = \psi$ or not.

5.26.4   my take on this in 3D for MC

I have established in fieldstone that in the general 3D case

$$
\mathcal{F}^{\mathrm{MC}} = \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sin\phi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta_{\mathrm{L}}(\boldsymbol{\tau}) - \frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}}(\boldsymbol{\tau})\sin\phi\right) - c\cos\phi
\tag{5.132}
$$

Note that since $p = -\mathcal{I}_1(\boldsymbol{\sigma})/3$ then we recover the usual '$p\sin\phi + c\cos\phi$'.
  Following Eq. (4) of the paper the plastic potential would be given by

$$
\mathcal{Q}^{\mathrm{MC}} = \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sin\psi + \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}\left(\cos\theta_{\mathrm{L}}(\boldsymbol{\tau}) - \frac{1}{\sqrt{3}}\sin\theta_{\mathrm{L}}(\boldsymbol{\tau})\sin\psi\right)
$$

The visco-plastic strain rate would then write

$$
\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma\langle\phi(\mathcal{F}^{\mathrm{MC}})\rangle\frac{\partial \mathcal{Q}^{\mathrm{MC}}}{\partial \boldsymbol{\sigma}}
$$

We have established that

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} &= C_1\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + C_2\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + C_3\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \\
&= C_1\mathbf{1} + C_2\boldsymbol{\tau} + C_3\left(\boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1}\right)
\end{aligned}
\tag{5.133}
$$

and in the case of the Mohr-Coulomb criterion:

$$
C_1^{\mathrm{MC}} = \frac{1}{3}\sin\phi
\tag{5.134}
$$

$$
C_2^{\mathrm{MC}} = \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\cos\theta_{\mathrm{L}}\left[(1 + 2\tan\theta_{\mathrm{L}}\tan 3\theta_{\mathrm{L}}) + \frac{1}{\sqrt{3}}\sin\phi(2\tan 3\theta_{\mathrm{L}} - \tan\theta_{\mathrm{L}})\right]
$$

$$
C_3^{\mathrm{MC}} = \frac{\sqrt{3}\sin\theta_{\mathrm{L}} + \sin\phi\cos\theta_{\mathrm{L}}}{2\mathcal{I}_2(\boldsymbol{\tau})\cos 3\theta_{\mathrm{L}}}
\tag{5.135}
$$

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma \langle \phi(\mathcal{F}) \rangle \left(C_1 \mathbf{1} + C_2 \boldsymbol{\tau} + C_3 \left(\boldsymbol{\tau} \cdot \boldsymbol{\tau} - \frac{2}{3} \mathcal{I}_2(\boldsymbol{\tau}) \mathbf{1} \right) \right)$$

Assuming brackets ok, and $\phi(x) = x^n$:

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \gamma (\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y)^n \left(C_1 \mathbf{1} + C_2 \boldsymbol{\tau} + C_3 \left(\boldsymbol{\tau} \cdot \boldsymbol{\tau} - \frac{2}{3} \mathcal{I}_2(\boldsymbol{\tau}) \mathbf{1} \right) \right)$$

Taking the deviatoric part of this:

$$\dot{\boldsymbol{\varepsilon}}^{vp,d} = \gamma (\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - Y)^n \left(C_2 \boldsymbol{\tau} + C_3 \left(\boldsymbol{\tau} \cdot \boldsymbol{\tau} - \frac{2}{3} \mathcal{I}_2(\boldsymbol{\tau}) \mathbf{1} \right) \right)$$

so I cannot find a <u>scalar</u> η such that

$$\dot{\boldsymbol{\varepsilon}}^{vp,d} = \frac{1}{2\eta} \boldsymbol{\tau}$$

I am STUCK here?!

5.26.5 Revisiting Lemiale et al (2008) and Spiegelman et al (2016)

<div align="right">perzyna_others.tex</div>

The authors postulate that the total strain rate is the sum of the viscous deformation and plastic deformation:

$$\dot{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}}^v + \dot{\boldsymbol{\varepsilon}}^p$$

Then

$$\boldsymbol{\sigma} = -p\mathbf{1} + 2\eta(\dot{\boldsymbol{\varepsilon}} - \dot{\boldsymbol{\varepsilon}}^p) \tag{5.136}$$

Immediately we see that they implicitly assume that the flow is incompressible. Upon yielding a flow rule is needed to specify the plastic behaviour. The plastic strain rate is written as

$$\dot{\boldsymbol{\varepsilon}}^p = \dot{\lambda} \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} \tag{5.137}$$

where $\dot{\lambda}$ is a scalar plastic flow rate and $\mathcal{Q}$ is the so-called plastic potential. Note that in Heeres, Suiker, and de Borst [591] the authors define $\dot{\lambda} = \langle \phi(x) \rangle / \eta$ so that the equation above is the Perzyna model. A classical choice for $\mathcal{Q}$, in conjunction with the incompressibility constraint, is:

$$\mathcal{Q} = \sqrt{\mathcal{I}_2(\boldsymbol{\tau})}$$

We notice that Eq. 5.137 is different (although obviously not unrelated) than the Perzyna approach above, although in the end they arrive at a similar expression as we did before for the von Mises case.

If we consider only the deviatoric part of the stress tensor in Eq. (5.136), we thus obtain[23]

$$\boldsymbol{\tau} = 2\eta(\dot{\boldsymbol{\varepsilon}} - \dot{\boldsymbol{\varepsilon}}^p) = 2\eta \left(\dot{\boldsymbol{\varepsilon}} - \dot{\lambda} \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} \right) = 2\eta \left(\dot{\boldsymbol{\varepsilon}} - \dot{\lambda} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right) = 2\eta \left(\dot{\boldsymbol{\varepsilon}} - \dot{\lambda} \frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \boldsymbol{\tau} \right)$$
$$\tag{5.138}$$

This equation can be written as

$$\left(1 + \dot{\lambda} \frac{\eta}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}} \right) \boldsymbol{\tau} = 2\eta \dot{\boldsymbol{\varepsilon}}$$

---

[23]do they assume varepsilon deviatoric too?

One can then take the square root of the second invariant of this equation:

$$\left(1 + \dot{\lambda}\frac{\eta}{\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\right)\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} = 2\eta\sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}})}$$

Then

$$\left(1 + \dot{\lambda}\frac{\eta}{\tau_e}\right)\tau_e = 2\eta\dot{\varepsilon}_e$$

so that

$$\dot{\lambda} = \frac{2\eta\dot{\varepsilon}_e - \tau_e}{\eta} = 2\dot{\varepsilon}_e - \frac{\tau_e}{\eta}$$

Finally we can insert this expression of $\dot{\lambda}$ in Eq. (5.138)

$$
\begin{aligned}
\boldsymbol{\tau} &= 2\eta\left(\dot{\boldsymbol{\varepsilon}} - \dot{\lambda}\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\boldsymbol{\tau}\right) \\
&= 2\eta\left(\dot{\boldsymbol{\varepsilon}} - (2\dot{\varepsilon}_e - \frac{\tau_e}{\eta})\frac{1}{2\sqrt{\mathcal{I}_2(\boldsymbol{\tau})}}\boldsymbol{\tau}\right) \\
&= 2\eta\left(\dot{\boldsymbol{\varepsilon}} - (2\dot{\varepsilon}_e - \frac{\tau_e}{\eta})\frac{1}{2\tau_e}\boldsymbol{\tau}\right) \\
&= 2\eta\dot{\boldsymbol{\varepsilon}} - 2\eta\dot{\varepsilon}_e\frac{1}{\tau_e}\boldsymbol{\tau} + \eta\frac{\tau_e}{\eta}\frac{1}{\tau_e}\boldsymbol{\tau} \\
&= 2\eta\dot{\boldsymbol{\varepsilon}} - 2\eta\dot{\varepsilon}_e\frac{1}{\tau_e}\boldsymbol{\tau} + \boldsymbol{\tau}
\end{aligned}
\tag{5.139}
$$

The term $\boldsymbol{\tau}$ is present on both sides of the equal sign so it cancels out and we are left with:

$$\mathbf{0} = 2\eta\dot{\boldsymbol{\varepsilon}} - 2\eta\dot{\varepsilon}_e\frac{1}{\tau_e}\boldsymbol{\tau}$$

or,

$$\boldsymbol{\tau} = \frac{\tau_e}{\dot{\varepsilon}_e}\dot{\boldsymbol{\varepsilon}}$$

On yield we have $\tau_e = Y(c,\phi)$ so in the end:

$$\boldsymbol{\tau} = 2\underbrace{\frac{1}{2}\frac{Y(c,\phi)}{\dot{\varepsilon}_e}}_{\eta_p}\dot{\boldsymbol{\varepsilon}}$$

That last step is poorly documented in the paper! This is a cumbersome exercise and it relies heavily on the choice of $\mathcal{Q} = \tau_e$.

Remark. *Lemiale et al. (2008) define $\bar{\tau} = \sqrt{\tau_{ij}\tau_{ij}/2}$ but define $\dot{\gamma} = \sqrt{2D_{ij}D_{ij}}$!*

Let us now turn to Spiegelman, May, and Wilson [1267] (2016). In Section 2.1.1 of this paper the authors follow the same path as above. They assume $\mathcal{Q} = \tau_e$ but justify their choice by stating that "The use of incompressible materials mandates that we use a plastic potential g which is not a function of the pressure p"

This is indeed very important in the context of our incompressible calculations in geodynamics.

They define the yield surface, $\mathcal{F}(\boldsymbol{\sigma})$ which is a scalar function defining the failure (yield) state of a material. Yield surfaces are assumed to be of the following form

$$\mathcal{F}(\boldsymbol{\sigma}) = \tau_e - Y(\boldsymbol{\sigma})$$

where Y is the yield criterion. The authors state that "it is common practice in geodynamics to define the plastic multiplier $\dot{\lambda}$ which exactly satisfies $\mathcal{F} = 0$, or equivalently $\tau_e = Y$ [810]". We see that it is then the same as the Lemiale *et al.* paper.

Chapter 6

The Finite Difference Method

6.1 Back to basics: what is a derivative?

Before we start with the basics of the Finite Difference method, we should quickly recall the definition of the derivative of a function.

> The derivative of a function $y = f(x)$ of a variable x is a measure of the rate at which the value y of the function changes with respect to the change of the variable x. It is called "the derivative of f with respect to x". If x and y are real numbers, and if the graph of f is plotted against x, the derivative is the <u>slope</u> of this graph at each point.

There are classically two standard notations:

$$\frac{df}{dx}(x) \qquad \text{and} \qquad f'(x) \tag{6.1}$$

The mathematical definition is[1]:

$$\boxed{f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{(x+h) - x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}} \tag{6.2}$$

i.e., how much does the function change between x and $x + h$, divided by the length h. On the following left plot the function is in green while the line joining the points $x, f(x)$ and $x+h, f(x+h)$ is shown in purple. On the right figure we see that when h becomes smaller and smaller this line does indeed get closer and closer to the real tangent line in brown.
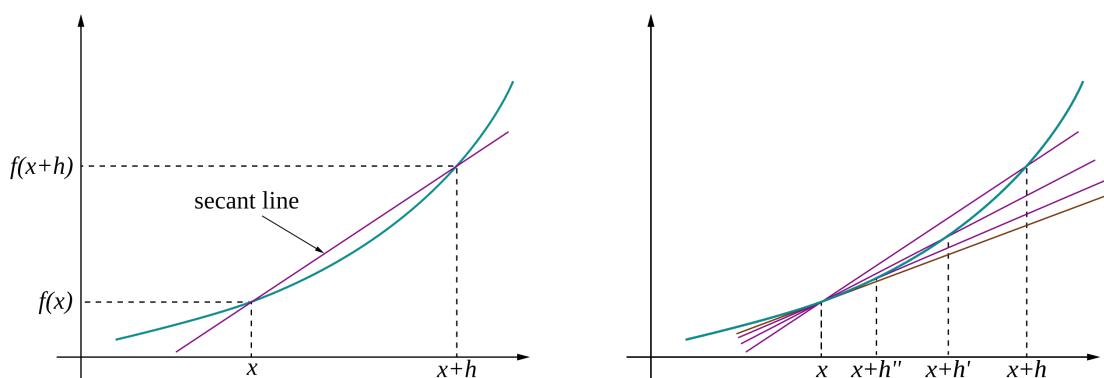
Left: The secant to curve $y = f(x)$ determined by points $(x, f(x))$ and $(x + h, f(x + h))$; Right: The tangent line as limit of secants $(h'' < h' < h)$. Taken from Wikipedia[2].

---

[1] if the limit exists

[2] https://en.wikipedia.org/wiki/Derivative

Also, one can rewrite the formula above as

$$f(x + h) \simeq f(x) + f'(x)h \tag{6.3}$$

which is in fact the beginning of the Taylor expansion of the function at position $x + h$:

$$f(x + h) \simeq f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + \dots \tag{6.4}$$

We will see in what follows that the Taylor expansion and the concept of derivation is central to the Finite Difference method.

6.2 Welcome to the discrete world

In mathematics and in physics we assume that space and time form a continuum, i.e. a segment can be divided in two indefinitely, or in other words one can 'zoom in' on a part of space or time as much as needed. However computers are binary machines with a finite amount of memory so they cannot represent a continuum (the transistors in a processor are either 'on' or 'off', nothing in between). As a consequence, in the context of solving PDEs describing physical phenomena, computers will only allow us to compute the solution at certain discrete locations and at certain discrete times.

Illustration of space discretisation. If the picture is cut up in a small number of cells covering it we obtain the picture on the left. If more cells are used, we obtain the picture in the middle. When a (very) large number of cells is used we finally see all details and it approaches a continuum.

6.3 FDM basics in 1d

In what follows we suppose that we have a function $f(x)$, which is continuous and differentiable over the range of interest. Let us also assume that we know the value $f(x_0)$ and all the derivatives at $x = x_0$.

6.3.1 First order derivatives

The forward Taylor-series expansion for $f(x_0 + h)$, away from the point x_0 by a small amount h is given by

$$f(x_0 + h) = f(x_0) + h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) + \dots + \frac{h^n}{n!}\frac{\partial^n f}{\partial x^n}(x_0) + \mathcal{O}(h^{n+1}) \tag{6.5}$$

We can substract $f(x_0)$ to each side of the equation and divide by h:

$$\frac{1}{h}\left[f(x_0 + h) - f(x_0)\right] = \frac{\partial f}{\partial x}(x_0) + \frac{h}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) + \ldots \tag{6.6}$$

and we can then express the first derivative of f as follows:

$$\frac{\partial f}{\partial x}(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2!}\frac{\partial^2 f}{\partial x^2}(x_0)\ldots \tag{6.7}$$

or, replacing the term in h by $\mathcal{O}(h)$:

$$\boxed{\frac{\partial f}{\partial x}(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} + \mathcal{O}(h)} \tag{6.8}$$

$\mathcal{O}(h)$ indicates that the full solution would require additional terms of order h, h^2, and so on. $\mathcal{O}$ is called the truncation error: if the distance h is made smaller and smaller, the (numerical approximation) error decreases $\propto h$ in this case.

Let us assume that the 1D domain on which a given Ordinary Differential Equation (ODE) or Partial Differential Equation (PDE) is to be solved has been discretised and let us zoom in on three consecutive points ($x_0 = x_i$ here):
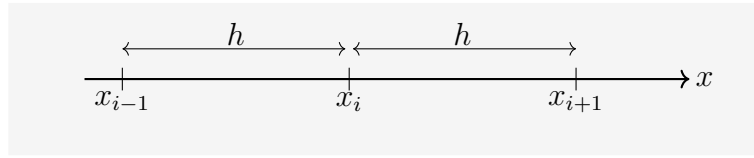
In the context of a discrete calculation on a set of discrete points x_i we can compute the first order derivative of f at point x_i as an approximation:

$$\boxed{\frac{\partial f}{\partial x}(x_i) = \frac{f_{i+1} - f_i}{h} + \mathcal{O}(h)} \qquad \text{(forward difference)} \tag{6.9}$$

where functions $f_i = f(x_i)$ are evaluated at discretely spaced x_i with $x_{i+1} = x_i + h$ (i.e. $h = x_{i+1} - x_i$), where the node spacing, or resolution, h is assumed constant. We also introduce the notation $f'_i = f'(x_i) = \frac{\partial f}{\partial x}(x_i)$.

The **forward FD derivative** as expressed above is called **first order accurate**, and this means that very small h is required for an accurate solution.

Example FDM-1: Before we go any further with the theory, let us look at a very simple example. Let us consider the Stokes equations in the absence of fluid motion (i.e. $\vec{v} = \vec{0}$). Then the strain rate tensor components are identically zero and the equation simply is

$$-\vec{\nabla}p + \rho\vec{g} = \vec{0}. \tag{6.10}$$

where we assume that 1) density is constant in space for simplicity and 2) the domain is infinite in the x-direction. The gravity vector is $\vec{g} = -g\vec{e}_y$ so that the above equation becomes:

$$-\frac{dp}{dy} - \rho g = 0 \tag{6.11}$$

This is a first-order ODE. It needs to be supplemented by a single boundary condition, which in this case constrains the pressure to be zero at the surface, i.e. $p(y = L_y) = 0$.
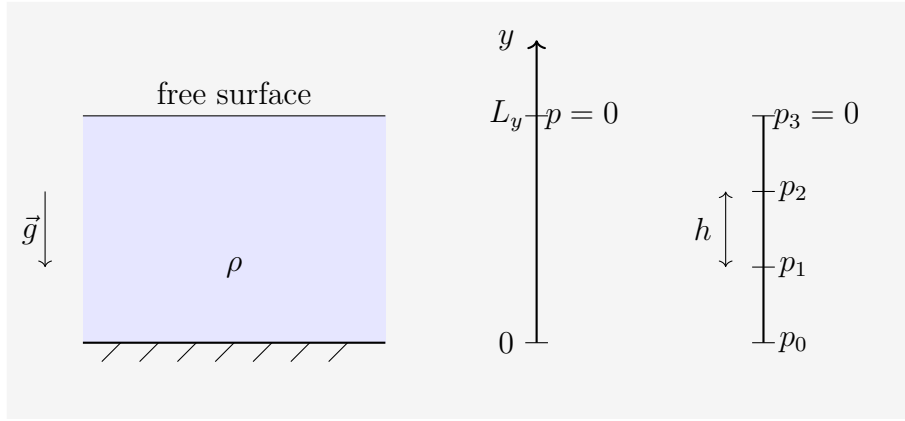
We can write the discretised ODE at node 2 (since we know $p_3 = 0$):

$$\frac{dp}{dy}(x_2) \simeq \frac{p_3 - p_2}{h} = -\rho g \tag{6.12}$$

or, $p_2 = \rho g h$. Having obtained p_2, we write the ODE at node 1:

$$\frac{dp}{dy}(x_1) \simeq \frac{p_2 - p_1}{h} = -\rho g \tag{6.13}$$

or, $p_1 = \rho g h + p_2 = \rho g 2h$. And finally we obtain as expected

$$p_0 = \rho g \, 3h = \rho g L_y. \tag{6.14}$$

This brings us to our first exercise:

We will now put the previous example into practice and write a python code which uses forward differences to compute the 1D pressure field inside the crust.

```python
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
Ly=30e3
nny=10
g=9.81
rho=3000
#Compute , the distance between nodes
h=
#declare/fill the y array
y=

#Set the pressure at the top to zero and work backwards to compute the pressure a

#Print the expected value at the bottom next to the computed value at y=0

plt.figure(figsize=(6,5)) #Size of figure in inches (or 'freedom units'), x,y
plt.title('Pressure Profile')
plt.plot(p,y,color='black')
plt.xlabel('Pressure (N/m2)')
plt.ylabel('Vertical Profile (m)')
plt.grid()
plt.savefig('PressureProfile.pdf')
```

We can also expand the Taylor series backward (i.e. looking 'left' of x_0)

$$f(x_0 - h) = f(x_0) - h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) - \dots \qquad (6.15)$$

The **backward FD derivative** then writes:

$$\boxed{\frac{\partial f}{\partial x}(x_i) = \frac{f_i - f_{i-1}}{h} + \mathcal{O}(h)} \qquad \text{(backward difference)} \qquad (6.16)$$

Alternatively, we can substract the backward formula from the forward one and divide by two; concretely, we start from

$$f(x_0 + h) = f(x_0) + h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) + \dots \qquad (6.17)$$

and substract the following from it

$$f(x_0 - h) = f(x_0) - h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) + \dots \qquad (6.18)$$

to obtain:

$$f(x_0 + h) - f(x_0 - h) = 2h\frac{\partial f}{\partial x}(x_0) + \mathcal{O}(h^3) \qquad (6.19)$$

or,

$$\frac{\partial f}{\partial x}(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + \mathcal{O}(h^2)$$

We see that the resulting **central difference** approximation is **second order accurate**. In the discrete world one then write

$$\boxed{\frac{\partial f}{\partial x}(x_i) = \frac{f_{i+1} - f_{i-1}}{2h} + \mathcal{O}(h^2)} \qquad \text{(central difference)} \qquad (6.20)$$

Simply put, the denominator is $2h$ because it is the distance between point x_{i-1} and x_{i+1}.
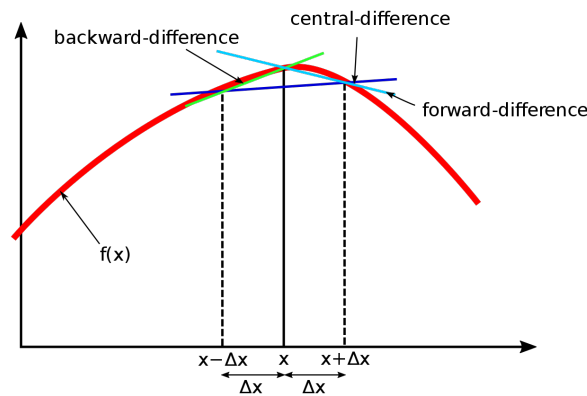
3 types of the finite difference method. Central gives the best approximation of the derivative. Taken from Wikipedia[3]

Can we do better than $\mathcal{O}(h^2)$? The answer is yes, and I list hereunder the formula:

$$\frac{\partial f}{\partial x}(x_i) = \frac{2f_{i+1} + 3f_i - 6f_{i-1} + f_{i-2}}{6h} + \mathcal{O}(h^3) \qquad \text{backward difference} \qquad (6.21)$$

$$\frac{\partial f}{\partial x}(x_i) = \frac{-f_{i+2} + 6f_{i+1} - 3f_i - 2f_{i-1}}{6h} + \mathcal{O}(h^3) \qquad \text{forward difference} \qquad (6.22)$$

[3]`https://en.wikipedia.org/wiki/Finite_difference`

$$\frac{\partial f}{\partial x}(x_i) = \frac{-f_{i+2} + 8f_{i+1} - 8f_{i-1} + f_{i-2}}{12h} + \mathcal{O}(h^4) \qquad \text{central difference} \tag{6.23}$$

Looking at these formula it is obvious that the cost of forming the derivative is larger than before (more multiplications, additions, ...) which translates to longer calculations and, in the case of implicit methods, much denser matrices.

Exercise FDM-1 (bonus)

Prove the formula above.

6.3.2 Second-order derivatives

Many PDEs contain second order derivatives (typically diffusion equations or wave equations) so we now turn to these and define $f_i'' = f''(x_i) = \frac{d^2 f}{dx^2}(x_i)$.

Second order derivative - forward   Let us define a function $g(x)$ such that $g = f'$. Then we have seen that the forward difference formula leads to write:

$$g_i' = \frac{g_{i+1} - g_i}{h} \tag{6.24}$$

On the one hand, we have $g_i' = g'(x_i) = f''(x_i) = f_i''$ and on the other hand

$$\frac{g_{i+1} - g_i}{h} = \frac{f_{i+1}' - f_i'}{h} \tag{6.25}$$

We can then use the forward derivative formula for f_{i+1}' and f_i' and obtain the following second order derivatives of f:

$$f_i'' = \frac{f_{i+1}' - f_i'}{h} = \frac{\frac{f_{i+2} - f_{i+1}}{h} - \frac{f_{i+1} - f_i}{h}}{h} = \frac{f_{i+2} - 2f_{i+1} + f_i}{h^2} \tag{6.26}$$

which is the **first order accurate**, **forward difference** approximation for second order derivatives at x_i. In order to compute $f''(x_i)$ we need the value of f at x_i but also at two other locations right of this location.

Second order derivative - backward   Likewise, we obtain the following formula when using the backward derivative twice:

$$f_i'' = \frac{f_i' - f_{i-1}'}{h} = \frac{\frac{f_i - f_{i-1}}{h} - \frac{f_{i-1} - f_{i-2}}{h}}{h} = \frac{f_i - 2f_{i-1} + f_{i-2}}{h^2} \tag{6.27}$$

This time we we need the value of f at x_i but also at two other locations left of this location.

Second order central   By adding the taylor expansions (with $+h$ and $-h$) a **second order accurate** approximation of the second derivative is obtained. We start from

$$
\begin{aligned}
f(x_0 + h) &= f(x_0) + h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) + \frac{h^3}{3!}\frac{\partial^3 f}{\partial x^3}(x_0) + \cdots + \frac{h^n}{n!}\frac{\partial^n f}{\partial x^n}(x_0) + \mathcal{O}(h^{n+1}) \\
f(x_0 - h) &= f(x_0) - h\frac{\partial f}{\partial x}(x_0) + \frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0) - \frac{h^3}{3!}\frac{\partial^3 f}{\partial x^3}(x_0) + \cdots + \frac{(-h)^n}{n!}\frac{\partial^n f}{\partial x^n}(x_0) + \mathcal{O}(h^{n+1})
\end{aligned}
$$

and we see that adding the first equation to the second yields

$$f(x_0+h)+f(x_0-h) = 2f(x_0)+\underbrace{h\frac{\partial f}{\partial x}(x_0) - h\frac{\partial f}{\partial x}(x_0)}_{=0}+2\frac{h^2}{2!}\frac{\partial^2 f}{\partial x^2}(x_0)+\underbrace{\frac{h^3}{3!}\frac{\partial^3 f}{\partial x^3}(x_0) - \frac{h^3}{3!}\frac{\partial^3 f}{\partial x^3}(x_0)}_{=0}+\mathcal{O}(h^4)$$

(6.28)

or,

$$\frac{\partial^2 f}{\partial x^2}(x_0) = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} + \mathcal{O}(h^2)$$

(6.29)

which translates into

$$\boxed{f_i'' = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} + \mathcal{O}(h^2)}$$ (second order central difference)   (6.30)

Note that this formula requires one value left and one value right of the point under consideration.

Another way to arrive at the same expression is to write the expansion at $x_0 \pm h/2$, i.e. at the (convenient, yet nonexistent) half points $i \pm 1/2$:
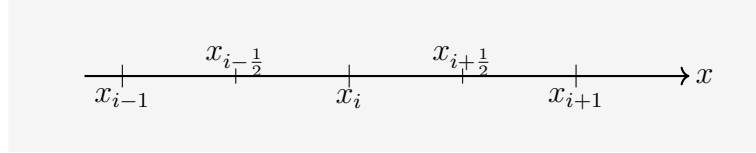
$$f_{i+1/2}' = \frac{f_{i+1} - f_i}{h} \qquad\qquad f_{i-1/2}' = \frac{f_i - f_{i-1}}{h}$$

(6.31)

$$f_i'' = \frac{f_{i+1/2}' - f_{i-1/2}'}{h} = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}$$

(6.32)

Note that derivatives of the form (see heat transport equation in Section 3.6)

$$\frac{\partial}{\partial x}\left(k\frac{\partial f}{\partial x}\right)$$

(6.33)

where k is a function of space, should be formed as follows

$$\frac{\partial}{\partial x}\left(k\frac{\partial f}{\partial x}\right)\bigg|_i = \frac{k_{i+1/2}\frac{f_{i+1}-f_i}{h} - k_{i-1/2}\frac{f_i-f_{i-1}}{h}}{h} + \mathcal{O}(h^2)$$

(6.34)

where $k_{i\pm1/2}$ is evaluated between the points to maintain the second order accuracy.

Remark. *If the heat conductivity k shows strong jumps from one grid point to another that are not aligned with the grid-nodes, most second-order methods will show first order accuracy at best.*

Can we do better than $\mathcal{O}(h^2)$? The answer is yes again:

$$f_i'' = \frac{-f_{i+2} + 16f_{i+1} - 30f_i + 16f_{i-1} - f_{i-2}}{12h^2} + \mathcal{O}(h^4)$$

(6.35)

Note that in some cases the fully forward or backward expressions need to be used on the boundary if the value of the function is not prescribed.

6.4 Solving the 1d diffusion equation

Consider the one-dimensional, transient (i.e. time-dependent) heat conduction equation without heat generating sources

$$\rho C_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) \tag{6.36}$$

where ρ is density, C_p heat capacity, k thermal conductivity, T temperature, x distance, and t time.

If the thermal conductivity, density and heat capacity are constant over the model domain, the equation can be simplified to a diffusion equation:

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} \tag{6.37}$$

where $\kappa = k/\rho C_p$ is the heat diffusivity.

We wish to solve this PDE in time and space (provided the appropriate boundary conditions have been given). The domain is $[0, L_x]$ and it is discretised by means of nnx points as depicted hereunder:
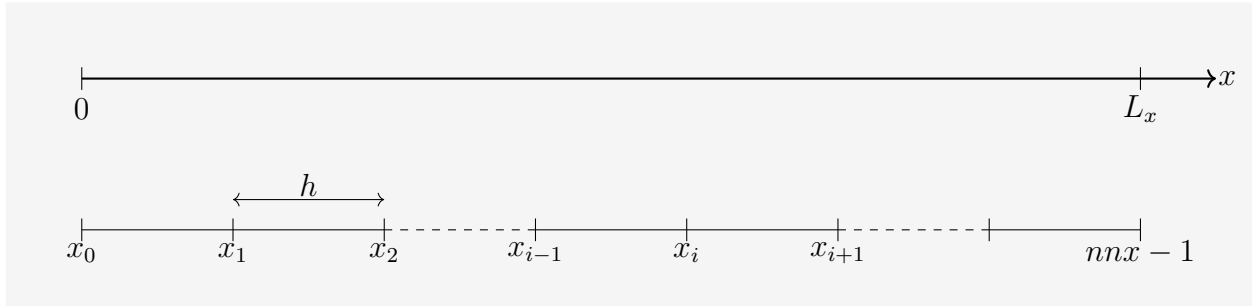
The derivative of temperature with regards to time can be approximated with a forward finite difference approximation *in time* as

$$\frac{\partial T}{\partial t} \simeq \frac{T_i^{n+1} - T_i^n}{t^{n+1} - t^n} = \frac{T_i^{n+1} - T_i^n}{\delta t} \tag{6.38}$$

where δt is the time step, i.e. the time between two consecutive measurements (the equivalent of h in space). In all that follows the subscript will always refer to space indices while the superscript will always refer to time indices. To be clear: n represents the current time step whereas $n+1$ represents the next time step.

Both n and i are integers; n varies from 0 to $nstep - 1$ (total number of time steps) and i varies from 0 to $nnx - 1$ (where nnx is the total number of grid points in x-direction)[4].

The spatial derivative is replaced by a central FD approximation expressed at t^n:

$$\frac{\partial^2 T}{\partial x^2} \simeq \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} \tag{6.39}$$

We then obtain

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = \kappa \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} \tag{6.40}$$

---

[4]I here adopt the zero-based numbering used in Python, see `https://en.wikipedia.org/wiki/Zero-based_numbering`

and finally

$$T_i^{n+1} = T_i^n + \delta t\, \kappa \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} \tag{6.41}$$

Because the temperature at the current time step n is known, we can compute the new temperature without solving any additional equations. Such a scheme is an **explicit** finite difference method and was made possible by the choice to evaluate the temporal derivative with forward differences.

Explicit scheme
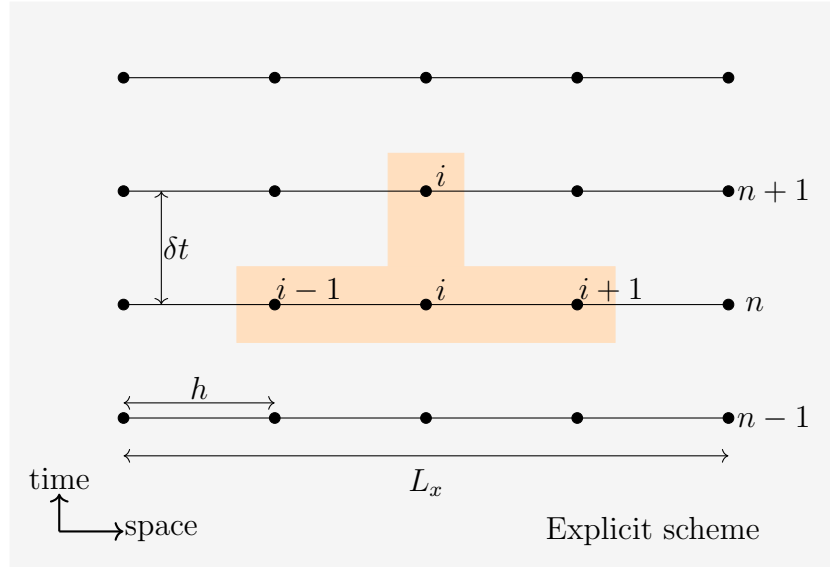
In order to solve the original PDE equation we need to

- prescribe an initial temperature field

- prescribe two boundary conditions

Such requirements hold also in the discrete world.

We know that this numerical scheme will converge to the exact solution for small h and δt because it has been shown to be consistent - that its discretization process can be reversed, through a Taylor series expansion, to recover the governing partial differential equation - and because it is stable for certain values of h and δt: any spontaneous perturbations in the solution (such as round-off error) will either be bounded or will decay.

Example FDM-2: let us prescribe an initial temperature field T_i^0 for $i = 0, nnx - 1$. For example:
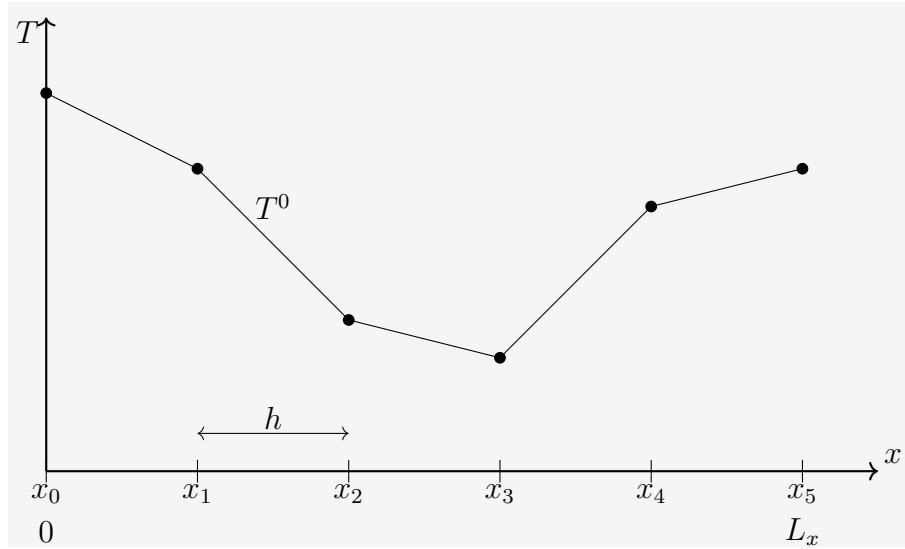
Then, we will be able to compute the new temperature of (for example) node 3 at time $t = 1 \cdot \delta t$ (i.e. T_3^1) with

$$T_3^1 = T_3^0 + \delta t \, \kappa \frac{T_4^0 - 2T_3^0 + T_2^0}{h^2} \tag{6.42}$$

Note that T_0 and T_5 cannot be computed by means of the above equation, which is not a problem because both these values are actually the prescribed boundary conditions.

The main drawback of the explicit approach is that stable solutions are obtained *only* when

$$0 < \frac{2\kappa\delta t}{h^2} \leq 1 \qquad \text{or,} \qquad \delta t \leq \frac{h^2}{2\kappa} \tag{6.43}$$

If this condition is not satisfied, the solution becomes unstable, starts to wildly oscillate and ultimately 'blows up'. We will observe this during the practicals.

The stability condition means that the maximum time step needs to be smaller than the time it takes for an anomaly to diffuse across the grid (nodal) spacing h. The explicit solution is an example of a conditionally stable method that only leads to well behaved solutions if a criterion like the one above is satisfied.

We are going to solve the 1D diffusion equation with the explicit method for the following physical setup: The domain is $L_x = 1$km long, it is maintained at a temperature $T = 100°$C at $x = 0$ and at a temperature $T = 200°$C at $x = L_x$. The initial temperature is $T(x, t = 0) = 123$ and $\kappa = 10^{-6}$. Time stepping will be carried out until steady state is reached.

```
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt


# Finish the following function which returns the analytical solution
# to the 1D steady-state diffusion problem.

def analytical_solution(x,Lx,Tleft,Tright):
    return ...

Lx=1e3
nnx=10
kappa=1e-6
nstep=2
Tleft=100
Tright=200
epsilon=0.1

h=
print('h=',h)
dt=
print('dt=',dt)


# Declare the array 'x' which will contain the coordinate of all
# nnx points #and then fill this array.
x=
```

We will need to store at least two temperature fields: the one we already know $\vec{T}^n$ (the initial temperature or a previously computed field at time n), and the one we want to compute $\vec{T}^{n+1}$ (We choose not to store all transient fields as we are only interested in the final steady state field). We therefore use only two arrays: $Told$ and $Tnew$, both of length nnx. Declare these arrays. Fill $Told$ with the value for $\vec{T}^0$.

```
Told=
Tnew=
```

At each timestep n compute from $Told$ the new temperature $Tnew$, using boundary conditions and the explicit scheme. Having done so, compute the average error between $Tnew$ and the real analytical solution. Stop the timestepping loop if this (absolute) error is smaller than ϵ

```
for n in range(0,nstep):

    for i in range (0,nnx):

        # compute Tnew[i]

    #compute error

    #print min/max
    print("timestep #",n,' min/max Tnew=',np.min(Tnew),np.max(Tnew))

    #transfer new solution into Told array
    Told[:]=Tnew[:]
```

What follows is an example of how to plot the temperature field $Tnew$.

```
plt.figure(figsize=(6,5))
plt.title('Steady state temperature')
plt.plot(x,Tnew,color='black')
plt.xlabel('x (m)')
plt.ylabel('T (K)')
plt.grid()
plt.savefig('Temperature.pdf')
```

An alternative approach is an **implicit** finite difference scheme, where the spatial derivatives of the Laplacian are evaluated (at least partially) at the new time step. We then use the backward difference for the time derivative:

$$\frac{\partial T}{\partial t} = \frac{T_i^n - T_i^{n-1}}{\delta t} \tag{6.44}$$

so that

$$\frac{T_i^n - T_i^{n-1}}{\delta t} = \kappa \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} \tag{6.45}$$

Note that this is often rewritten as follows (i.e. one time step further, $n \to n+1$):

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = \kappa \frac{T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1}}{h^2} \tag{6.46}$$

It is a fully implicit scheme where the time derivative is taken backward. Let us define the dimensionless parameter s as follows:

$$s = \frac{\kappa \, \delta t}{h^2} \tag{6.47}$$

The previous equation can be rearranged as follows:

$$-s\,T_{i+1}^{n+1} + (1 + 2s)\,T_i^{n+1} - s\,T_{i-1}^{n+1} = T_i^n \qquad (6.48)$$

Implicit scheme
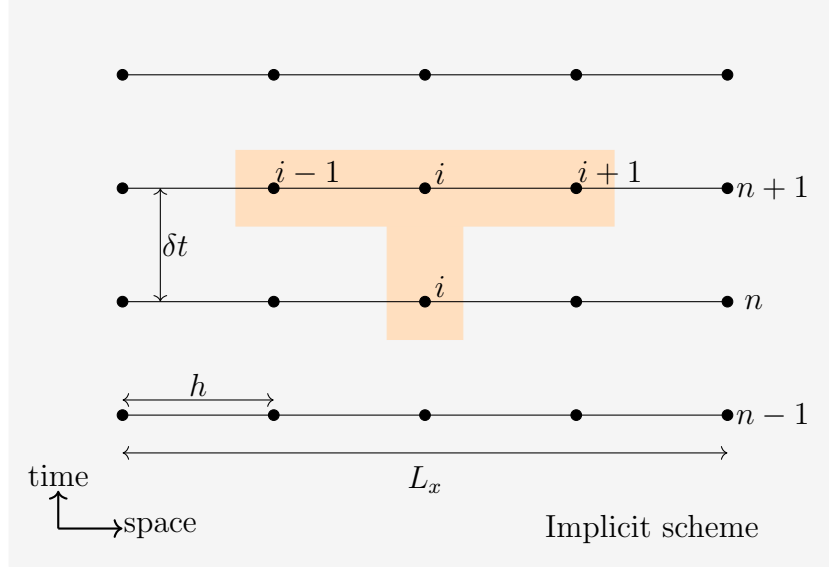
Note that in this case we no longer have an explicit relationship for T_{i-1}^{n+1}, T_i^{n+1} and T_{i+1}^{n+1}. Instead, we have to solve a linear system of equations, which is discussed further below.

The main advantage of implicit methods is that there are no restrictions on the time step, the fully implicit scheme is **unconditionally stable**. This does not mean that it is accurate as stability and accuracy are two different things. Taking large time steps may result in an inaccurate solution for features with small spatial scales.

For any application, it is therefore always a good idea to check the results by decreasing the time step until the solution does not change anymore (this is called a convergence check), and to ensure the method can deal with small and large scale features robustly at the same time.

Example FDM-3: Once again let us look at things with a very concrete approach. Let us discretise the domain of length L_x with 6 cells, i.e. $i = 0, \ldots 6$ ($nnx = 7$). We also prescribe the following boundary conditions (remember it is a 2nd order derivative in space, so we need two of them): $T(x = 0) = T_0 = 0$ and $T(x = L_x) = T_6 = 100$ (we assume that they do not change with time for simplicity). Finally we assume that we know T_i^0 for all i and we wish to compute T_i^1.

We then have:

$$
\begin{aligned}
T_0^1 &= 0 \\
-sT_2^1 + (1+2s)T_1^1 - sT_0^1 &= T_1^0 \\
-sT_3^1 + (1+2s)T_2^1 - sT_1^1 &= T_2^0 \\
-sT_4^1 + (1+2s)T_3^1 - sT_2^1 &= T_3^0 \\
-sT_5^1 + (1+2s)T_4^1 - sT_3^1 &= T_4^0 \\
-sT_6^1 + (1+2s)T_5^1 - sT_4^1 &= T_5^0 \\
T_6^1 &= 100
\end{aligned}
\tag{6.49}
$$

or,

$$
\underbrace{\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-s & 1+2s & -s & 0 & 0 & 0 & 0 \\
0 & -s & 1+2s & -s & 0 & 0 & 0 \\
0 & 0 & -s & 1+2s & -s & 0 & 0 \\
0 & 0 & 0 & -s & 1+2s & -s & 0 \\
0 & 0 & 0 & 0 & -s & 1+2s & -s \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}}_{\boldsymbol{A}}
\cdot
\underbrace{\begin{pmatrix}
T_0^1 \\ T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \\ T_5^1 \\ T_6^1
\end{pmatrix}}_{\vec{T}}
=
\underbrace{\begin{pmatrix}
0 \\ T_1^0 \\ T_2^0 \\ T_3^0 \\ T_4^0 \\ T_5^0 \\ 100
\end{pmatrix}}_{\vec{b}}
$$

As opposed to the explicit approach we must solve a linear system which size is given by the total number of nodes/points nnx in order to compute a new temperature field.

In summary, an implicit method requires us to solve $\boldsymbol{A} \cdot \vec{T} = \vec{b}$ with

- $\boldsymbol{A}$ is a $nnx \times nnx$ sparse matrix (i.e. mostly empty),

- $\vec{b}$ is a known vector of size nnx (often called the 'right-hand side', or rhs)

- $\vec{T}$ the vector of unknowns.

A word about solvers There are two main approaches to solving such a linear system: one can use a direct approach or an iterative approach. In a nutshell, a direct solver will 'manipulate' the matrix lines and columns so as to arrive at the solution (like you would do on paper yourself for a small system). A simple example of such an approach is the technique of elimination of variables of the following example.

Example FDM-4: Consider the following system:

$$\begin{pmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \\ 8 \end{pmatrix} \tag{6.50}$$

which is of course equivalent to

$$\begin{aligned} x + 3y - 2z &= 5 \\ 3x + 5y + 6z &= 7 \\ 2x + 4y + 3z &= 8 \end{aligned} \tag{6.51}$$

Solving the first equation for x gives $x = 5 + 2z - 3y$, and plugging this into the second and third equation yields (or take the second line of the matrix and remove 3 times the first line from it, etc ...)

$$\begin{pmatrix} 1 & 3 & -2 \\ 0 & -4 & 12 \\ 0 & -2 & 7 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ -8 \\ -2 \end{pmatrix} \tag{6.52}$$

Solving the second line for y yields $y = 2 + 3z$, and plugging this into the second equation yields $z = 2$. We now have:

$$\begin{pmatrix} 1 & 3 & -2 \\ 0 & -4 & 12 \\ 0 & 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ -8 \\ 4 \end{pmatrix} \tag{6.53}$$

Substituting $z = 2$ into the second equation gives $y = 8$, and substituting $z = 2$ and $y = 8$ into the first equation yields $x = -15$. Therefore, the solution set is the single point $(x, y, z) = (-15, 8, 2)$.

Taken from https://en.wikipedia.org/wiki/System_of_linear_equations

---

This example is of course very naive and direct solvers often come in the form of very large numerical libraries which have been highly optimised to take advantage of the sparsity of the matrix in order to arrive at the solution in the lowest number of operations possible. In reality techniques such as LU decomposition or Cholesky decomposition are used.

Iterative solvers on the other hand compute the solution of the system by first postulating an initial guess for the solution and then by improving this guess iteratively until the termination criterion is met (see Section 16.28). There are two classes of iteratives methods in this context: stationary iterative methods (e.g. Jacobi, Gauss-Seidel, SSOR) and Krylov subspace methods (e.g. CG, GMRES, BiCG). At this stage things get real complicated and the details of iterative solvers are vastly out of the scope of this course[5] (see for instance the book by Saad [1166]).

---

[5] https://en.wikipedia.org/wiki/Iterative_method

Example FDM-5: the stationary Jacobi method. The matrix $\boldsymbol{A}$ is decomposed as follows:

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U} \tag{6.54}$$

where $\boldsymbol{D}$ is the diagonal of the matrix $\boldsymbol{A}$, $\boldsymbol{L}$ is the strict lower triangular part of $\boldsymbol{A}$ and $\boldsymbol{U}$ is the strict upper triangular part of $\boldsymbol{A}$. The iterative method is defined by:

$$\boldsymbol{D} \cdot \vec{T}^{k+1} = -(\boldsymbol{L} + \boldsymbol{U}) \cdot \vec{T}^k + \vec{b} \qquad k = 0, 1, \ldots \tag{6.55}$$

where $\vec{T}^0$ is the initial guess (often taken to be zero). Note that the superscript denotes the iteration number and has nothing to do with the time step in this context. This method is trivial to implement since the linear system on the left side of the equal sign involves a diagonal matrix. This can also be written

$$T_i^{k+1} = \frac{1}{A_{ii}} \left(b_i - \sum_{j \neq i} A_{ij} T_j^k \right) \qquad i = 1, 2, \ldots nnx \tag{6.56}$$

Looking at the previous example, we have

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{L} + \boldsymbol{U} = \begin{pmatrix} 0 & 3 & -2 \\ 3 & 0 & 6 \\ 2 & 4 & 0 \end{pmatrix} \tag{6.57}$$

We then start with the guess $\vec{T}^0 = \vec{0}$, so that for $k = 0$:

$$\vec{T}^1 = \boldsymbol{D}^{-1} \cdot \vec{b} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \\ 8 \end{pmatrix} = \begin{pmatrix} 5 \\ 7/5 \\ 8/3 \end{pmatrix} \tag{6.58}$$

and then we obtain $\vec{T}^2$ by solving

$$
\begin{aligned}
\vec{T}^2 &= \boldsymbol{D}^{-1} \cdot \left[-(\boldsymbol{L} + \boldsymbol{U}) \cdot \vec{T}^1 + \vec{b} \right] \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \cdot \left[-\begin{pmatrix} 0 & 3 & -2 \\ 3 & 0 & 6 \\ 2 & 4 & 0 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7/5 \\ 8/3 \end{pmatrix} + \begin{pmatrix} 5 \\ 7 \\ 8 \end{pmatrix} \right] \\
&= \ldots
\end{aligned}
\tag{6.59}
$$

We keep iterating until two consecutively obtained temperature vectors are nearly identical, or,

$$\|\vec{T}^{k+1} - \vec{T}^k\| < \epsilon \tag{6.60}$$

where ϵ is a carefully chosen small enough number.

A sufficient (but not necessary) condition for the method to converge is that the matrix $\boldsymbol{A}$ is strictly or irreducibly diagonally dominant[a]. Strict row diagonal dominance means that for each row, the absolute value of the diagonal term is greater than the sum of absolute values of other terms $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$. Also, this algorithm will fail if one or more diagonal terms of $\boldsymbol{A}$ is nul.

---

[a]https://en.wikipedia.org/wiki/Diagonally_dominant_matrix

Exercise FDM-3

Implement example FDM-5 from scratch in a new code. Write a separate function for the Jacobi solver. What do you observe ? why does it explode? Multiply all diagonal values by 10 and re-run it. What do you observe now? Implement the Gauss-Seidel method. Which of the two methods converges the fastest?

$\rightarrow$`https://en.wikipedia.org/wiki/Iterative_method`.

Instructions: for each solver type create a dedicated function (e.g. `jacobi_solver`) which takes four arguments: the matrix $\boldsymbol{A}$, the vector $\vec{b}$, the guess vector and the desired tolerance set by the user.

Finally, looking at

$$-s\,T_{i+1}^{n+1} + (1+2s)\,T_i^{n+1} - s\,T_{i-1}^{n+1} = T_i^n \tag{6.61}$$

and dividing by $-s$ and letting $\delta t \rightarrow \infty$ (so $s \rightarrow \infty$ also), we obtain:

$$T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1} = 0 \tag{6.62}$$

which is a central difference approximation of the steady state solution

$$\frac{\partial^2 T}{\partial x^2} = 0 \tag{6.63}$$

Therefore, the fully implicit scheme will always yield the right equilibrium solution but may not capture small scale, transient features.

Exercise FDM-4

This is exactly the same exercise as Exercise 2 but we are going to solve the 1D diffusion equation with the implicit method this time. First use a solver from scipy to solve the system, then implement your own Jacobi solver. Note that the Jacobi solver must be implemented as a function which is to be called inside the time loop.

In your code use an if statement which allows to choose between explicit and implicit, and another if statement which allows to choose between scipy solver and iterative solver.

Implement the SOR method and the Conjugate Gradient method in separate functions. Make sure they all return the correct solution. Compare the solution and the path to convergence as returned by Jacobi, Gauss-Seidel and SOR for the same tolerance ϵ.

$\rightarrow$`https://en.wikipedia.org/wiki/Iterative_method`

$\rightarrow$`https://en.wikipedia.org/wiki/Conjugate_gradient_method`

Crank-Nicolson scheme It turns out that this fully implicit method is second order accurate in space but only first order accurate in time, i.e. the error goes as $\mathcal{O}(h^2, \delta t)$.

It is possible to write down a scheme which is second order accurate both in time and in space (i.e. $\mathcal{O}(h^2, \delta t^2)$), e.g. the Crank-Nicolson [6] scheme which remains unconditionally stable.

The Crank-Nicolson method is the time analog of central spatial differences and is given by

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = \kappa \frac{1}{2} \left[\underbrace{\frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2}}_{\text{at time } n} + \underbrace{\frac{T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1}}{h^2}}_{\text{at time } n+1} \right] \tag{6.64}$$

We now define $s = \kappa \delta t / 2h^2$ so that the equation above can be rearranged as follows :

$$\boxed{-s\, T_{i+1}^{n+1} + (1+2s)\, T_i^{n+1} - s\, T_{i-1}^{n+1} = s\, T_{i+1}^n + (1-2s)\, T_i^n + s\, T_{i-1}^n} \tag{6.65}$$

Any partially implicit method is more complicated to compute as we need to infer the future solution at time $n + 1$ by solution (inversion) of a system of linear equations based on the known solution at time n.

Implicit scheme
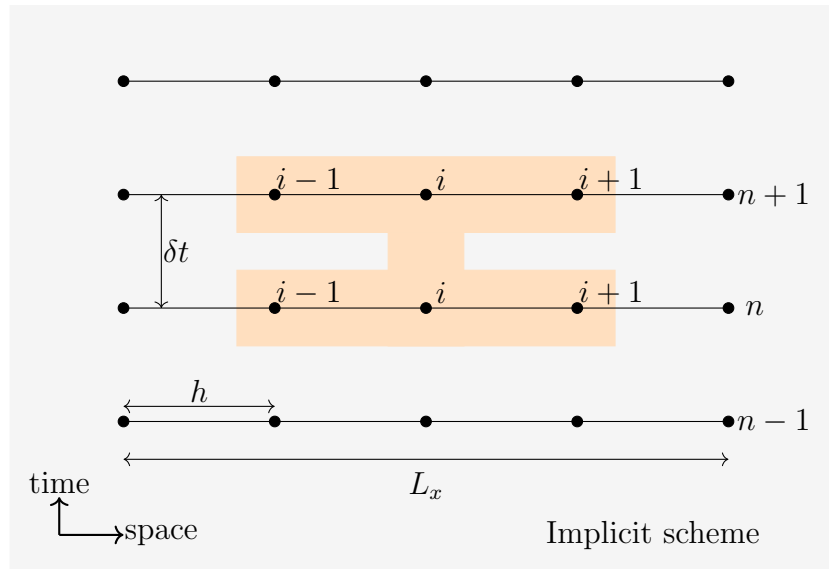
Exercise FDM-5

Modify the code of Exercise FDM-4 to implement the Crank-Nicolson method.

---

[6]The method was developed by John Crank and Phyllis Nicolson in the mid 20th century. `https://en.wikipedia.org/wiki/Crank-Nicolson_method`

Exercise FDM-6

Modify the code of Exercise FDM-5 to implement the more accurate discretisation of the second order derivative on the interior nodes (i.e. replace Eq. (6.30) by Eq. (6.35) for nodes $2, 3, 4, ... nnx - 3$, while you keep using Eq. (6.30) for nodes 1 and $nnx - 2$, and you prescribe temperature on nodes 0 and $nnx - 1$.

6.5 Solving the 1d advection equation

The 1d hyperbolic advection equation is:

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = 0 \tag{6.66}$$

or simply

$$\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} = 0 \tag{6.67}$$

We have seen how to deal with the time derivative (explicit, implicit) and with the first order space derivative (forward, backward or central). Let us consider the FTCS scheme (Forward in Time, Central in Space) expressed at node i:

$$\frac{T_i^{n+1} - T_i^n}{\delta t} + u_i \frac{T_{i+1}^n - T_{i-1}^n}{2h} = 0$$

or,

$$T_i^{n+1} = T_i^n - \frac{u_i \delta t}{2h}(T_{i+1}{}^n - T_{i-1}{}^n)$$

Note that although the velocity u is prescribed, it can vary in space, hence the subscript i. We here assume that it is time-independent for simplicity.

There is however a major problem: the FTCS method is in this case **unconditionally un**stable (see Section 6.2.1 of Hoffmann and Chiang [615], section 4.3.1 of Pelletier [1046]), i.e., it blows up for any δt. The instability is related to the fact that this scheme produces negative diffusion, which is numerically unstable.

The Crank-Nicolson implicit scheme for solving the diffusion equation can be adapted to solve the advection equation:

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = -u_i \frac{1}{2} \left(\frac{T_{i+1}^n - T_{i-1}^n}{2h} + \frac{T_{i+1}^{n+1} - T_{i-1}^{n+1}}{2h} \right)$$

or,

$$T_i^{n+1} + \frac{u_i \delta t}{4h}(T_{i+1}^{n+1} - T_{i-1}^{n+1}) = T_i^n - \frac{u \delta t}{4h}(T_{i+1}^n - T_{i-1}^n)$$

which then makes the method implicit.

We could also consider the FTFS method:

$$\frac{T_i^{n+1} - T_i^n}{\delta t} + u_i \frac{T_{i+1}^n - T_i^n}{h} = 0$$

190

but it is also **unconditionally un**stable (see Section 6.2.1 of [615]).

We will now look at to methods which alleviate this problem:

- The Lax-Friedrichs method[7] consists of replacing the $T_i{}^n$ in the time derivative term with $(T_{i+1}^n + T_{i-1}^n)/2$ (see for instance Section 4.3.1 of [1046] in the context of surface processes). The resulting equation is

$$\frac{T_i{}^{n+1} - (T_{i+1}^n + T_{i-1}^n)/2}{\delta t} = -u_i \frac{T_{i+1}^n - T_{i-1}^n}{2h}$$

or,

$$T_i{}^{n+1} = \frac{1}{2}(T_{i+1}^n + T_{i-1}^n) - \frac{u_i \delta t}{h}\frac{1}{2}(T_{i+1}^n - T_{i-1}^n)$$

von Neumann stability analysis indicates that this method is stable when $C = u\delta t/h \le 1$ where C is the Courant number.

- In the Streamline upwind method the spatial finite difference scheme depends on the sign of the velocity (assumed to be constant in space in what follows):

$$\frac{T_i{}^{n+1} - T_i{}^n}{\delta t} = \begin{cases} -u\frac{T_i^n - T_{i-1}^n}{h} & \text{if} \quad u > 0 \\[2mm] -u\frac{T_{i+1}^n - T_i^n}{h} & \text{if} \quad u < 0 \end{cases}$$

Using

$$\begin{aligned} u^+ &= \frac{1}{2}(u + |u|) \\ u^- &= \frac{1}{2}(u - |u|) \end{aligned} \tag{6.68}$$

i.e. if $u > 0$ then $u^+ = u$ and $u^- = 0$, if $u < 0$ then $u^+ = 0$ and $u^- = u$, so the upwind scheme can be written

$$\begin{aligned} 0 &= \frac{T_i{}^{n+1} - T_i{}^n}{\delta t} + \frac{\delta t}{h}\left[u^+(T_i^n - T_{i-1}^n) + u^-(T_{i+1}^n - T_i^n)\right] \\ &= \frac{T_i{}^{n+1} - T_i{}^n}{\delta t} + \frac{\delta t}{h}\left[\frac{1}{2}(u + |u|)(T_i^n - T_{i-1}^n) + \frac{1}{2}(u - |u|)(T_{i+1}^n - T_i^n)\right] \\ &= \frac{T_i{}^{n+1} - T_i{}^n}{\delta t} + \frac{u}{2h}(T_{i+1}{}^n - T_{i-1}{}^n) + \frac{|u|h}{2}\frac{-T_{i+1}{}^n + 2T_i{}^n - T_{i-1}{}^n}{h^2} \end{aligned} \tag{6.69}$$

or,

$$T_i{}^{n+1} = T_i{}^n - \frac{u\delta t}{2h}(T_{i+1}{}^n - T_{i-1}{}^n) + \frac{|u|h\delta t}{2}\frac{T_{i+1}{}^n - 2T_i{}^n + T_{i-1}{}^n}{h^2}$$

One can see that the 2nd term on RHS is the advection term obtained with the FTCS approach and the 3rd term has a form of diffusion.

Apparently, the 'diffusion term' included in the upwind scheme stabilizes the upwind scheme – it is achieved by damping the otherwise growing short waves. The included 'diffusion term' also introduces excessively damping to the short waves, as seen earlier. One possible remedy is to attempt to remove this excessive diffusion through one or several corrective steps. This is exactly what is done in Smolarkiewicz (1983, 1984) scheme, which is rather popular in the field of meteorology.

This method is stable when $C = u\delta t/h \le 1$. We will encounter Streamline Upwind again in the context of the Finite element method.

---

[7]Named after Peter Lax and Kurt O. Friedrichs. `https://en.wikipedia.org/wiki/Lax-Friedrichs_method`

- Lax-Wendroff method[8] [615] is second-order accurate in both space and time. This method is an example of explicit time integration where the function that defines the governing equation is evaluated at the current time.

$$T_i^{n+1} = T_i^n - \frac{u\delta t}{2h}(T_{i+1}{}^n - T_{i-1}{}^n) + \frac{u^2 dt^2}{2h^2}(T_{i+1}{}^n - 2T_i{}^n + T_{i-1}{}^n)$$

In [126] we find: " The Lax-Wendroff scheme has found considerable use in various applications; yet it is still somewhat diffusive and does not preserve sharp gradients as they are advected (see Rood, 1987, for an example). Another problem is what to do at the base of the grid (i=n). In an advective problem there is no boundary condition here; so if we apply the Lax-Wendroff at this point, it will contain the nonexistent concen-tration Bi+1 that cannot be eliminated with a boundary condition. One can apply [forward in time and backward in space equation] at this point and hope it does not perturb the scheme."

- modified Crank-Nicolson as shown in Becker and Kaus [71]:

$$\left[\frac{1}{6} - \frac{1}{4}\frac{u\delta t}{h}\right] T_{i-1}^{n+1} + \left[1 - \frac{1}{3}\right] T_i^{n+1} + \left[\frac{1}{6} + \frac{1}{4}\frac{u\delta t}{h}\right] T_{i+1}^{n+1}$$
$$= \left[\frac{1}{6} - \frac{1}{4}\frac{u\delta t}{h}\right] T_{i-1}^n + \left[1 - \frac{1}{3}\right] T_i^{n1} + \left[\frac{1}{6} + \frac{1}{4}\frac{u\delta t}{h}\right] T_{i+1}^n \tag{6.70}$$

These are not the only possibilities, see for instance the leapfrog method or the MacCormack method[9] (well suited for nonlinear equations). In p.346 of [126] we find: " The Leapfrog formulas are second order in both time and space, and they have no numerical dispersion. Nevertheless, they are not perfect because they generate spurious oscillations when they encounter sharp gradients, termed computational modes. These can grow and come to dominate the solution. Furthermore, the Leapfrog scheme inapplicable at the first time step; so a different start-up scheme needs to be used to generate the first set of data. This will introduce some small amount (we hope) of the bias in the result. In addition, as with the Lax-Wendroff, leapfrogging has a problem at the final point of the grid. Accepting that no classic scheme is perfect, Press et al. (1992) still prefer the Leapfrog scheme, at least when sharp gradients (shocks) are not involved. "

---

[8]Named after Peter Lax and Burton Wendroff https://en.wikipedia.org/wiki/Lax-Wendroff_method
[9]https://en.wikipedia.org/wiki/MacCormack_method

Let us consider the domain $[0, 1]$. The temperature field at $t = 0$ is given by $T = 1$ for $x < 0.25$ and $T = 0$ otherwise. The prescribed velocity is $u = 1$ and we set $nnx = 51$. Boundary conditions are $T = 1$ at $x = 0$ and $T = 0$ at $x = 1$.
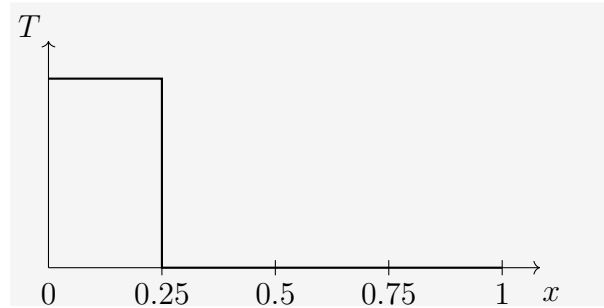
Program the above FTCS method. Run the model for 250 time steps with $\delta t = 0.002$. Program the Lax-Friedrichs method by modifying the previous code.

Bonus: Program the upwind method and/or the Crank-Nicolson method.

6.6 Solving the 1d advection diffusion equation

The 1d advection-diffusion equation is:

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = k \frac{\partial^2 T}{\partial x^2} + H' \tag{6.71}$$

or simply

$$\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} = \kappa \frac{\partial^2 T}{\partial x^2} + H \tag{6.72}$$

where $H = H'/\rho C_p$ is a source term that does not depend on time. We have seen how to deal with the time derivative (explicit, implicit) and with the first order space derivative (forward, backward or central). Let us consider the FTCS scheme (Forward in Time, Central in Space).

$$\frac{T_i^{n+1} - T_i^n}{\delta t} + u_i \frac{T_{i+1}^n - T_{i-1}^n}{2h} = \kappa \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} + H_i$$

Likewise we can consider the BTCS scheme:

$$\frac{T_i^{n+1} - T_i^n}{\delta t} + u_i \frac{T_{i+1}^{n+1} - T_{i-1}^{n+1}}{2h} = \kappa \frac{T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1}}{h^2} + H_i$$

which makes the method implicit. Or, using a Crank-Nicolson approach:

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = \frac{1}{2} \left(-u_i \frac{T^{n+1} n_{i+1} - T_{i-1}^{n+1}}{2h} + \kappa \frac{T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1}}{h^2} - u_i \frac{T_{i+1}^n - T_{i-1}^n}{2h} + \kappa \frac{T_{i+1}^n - 2T_i^n + T_{i-1}^n}{h^2} \right) + H$$

As we have seen before, the advection term is a source of problems and multiple methods have been designed to stabilise the pure advection equation. The Fiadeiro & Veronis method [419, 1470], works as follows for the (steady state) advection-diffusion equation (see also p315 of Boudreau [126]):

$$\kappa \frac{T_{i+1} - 2T_i + T_{i-1}}{h^2} - u \frac{(1-\sigma)T_{i+1} + 2\sigma T_i - (1+\sigma)T_{i-1}}{2h} = 0$$

which is a blend of backward (upstream) and central differences. The amount of blending is dictated by the value of the parameter σ, defined as

$$\sigma = \coth \frac{uh}{2\kappa} - \frac{2\kappa}{uh} = \coth \mathsf{Pe} - \frac{1}{\mathsf{Pe}}$$

The parameter σ has the property that $\sigma \to 0$ when $\mathsf{Pe} \to 0$ and $\sigma \to 1$ when $\mathsf{Pe} \to \infty$. The equation above can also be rewritten:

$$\kappa \frac{T_{i+1} - 2T_i + T_{i-1}}{h^2} - u \frac{1 T_{i+1} - T_{i-1}}{2h} = 0 + \frac{u\sigma h}{2} \frac{T_{i+1} - 2T_i + T_{i-1}}{h^2} = 0$$

which makes the action of this stabilisation term more obvious: it is a diffusion term whose diffusion coefficient goes away when $h \to 0$.

Thus, if $\sigma = 0$ (diffusion dominated), then pure central differencing is obtained, and if $\sigma = 1$ (advection dominated), then pure backward differencing results. Interestingly enough, this blended or weighted scheme is second-order accurate even as it switches to backward differencing [419]. Note that it is essentially the known streamline upwind method used in FEM or FDM (see Section 6.5). Also, when $\kappa \to 0$ (no physical diffusion at all) then σ tends to ± 1, and more precisely: $\sigma = sign(u)$, which allows for a simple and elegant implementation.

Adding a FT time derivative term, the equation above can be rewritten (a source term has been added):

$$\frac{T_i^{n+1} - T_i^n}{\delta t} = \kappa \frac{T_{i+1}^n - 2T_i + T_{i-1}^n}{h^2} - u \frac{(1-\sigma)T_{i+1}^n + 2\sigma T_i^n - (1+\sigma)T_{i-1}^n}{2h} + H$$

A simple example Let us consider the domain $0 < x < 1$. Let us prescribe $T(x = 0) = 0$ and $T(x = 1) = 0$ and we want to solve the steady state advection diffusion equation (with a source term):

$$u\frac{\partial T}{\partial x} = \kappa\frac{\partial^2 T}{\partial x^2} + 1 \tag{6.73}$$

The solution is then

$$T(x) = x - \frac{\exp(-(1-x)/\kappa) - \exp(-1/\kappa)}{1 - \exp(-1/\kappa)} \tag{6.74}$$

In this case, if $u \to 0$ then the equation becomes a diffusion equation and the solution is $T(x) = 0$. If u is large, the field is advected to the right but the right boundary condition $T(x = 1) = 0$ poses a problem which results in a very steep slope and potentially numerical issues too for the FTCS method, but not for the FTFV method.
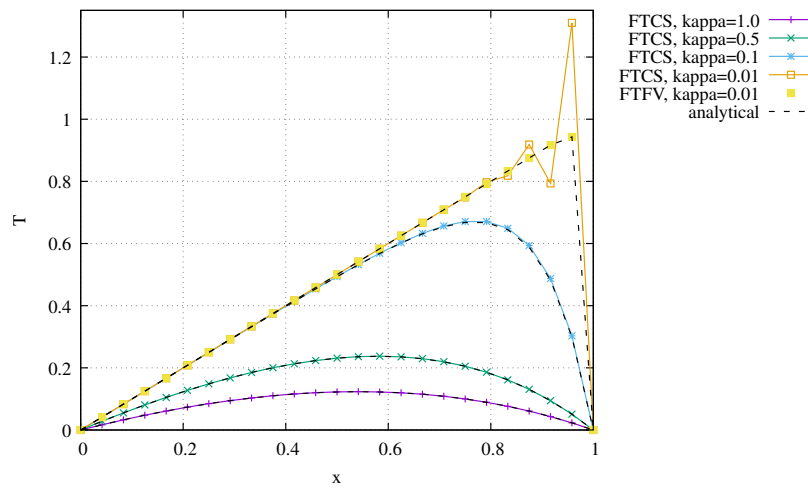
Results obtained with FTCS method. nnx=25, $u = 1$, $\delta t = 10^{-4}$, nstep=25000. FTFV stands for 'forward in time, Fiadeiro & Veronis'

A very similar example is implemented in STONE 65 (albeit with FEM).

Exercise FDM-8

Prove that Eq. (6.74) is indeed solution of Eq. (6.73). Code the FTCS and FTFV methods and reproduce the figure above (i.e. only forward in time/explicit methods).

6.7   FDM basics in 2d

In a 2D Cartesian domain overlain by a $nnx \times nny$ grid, the spacing between nodes in the x and y direction is h_x and h_y respectively.
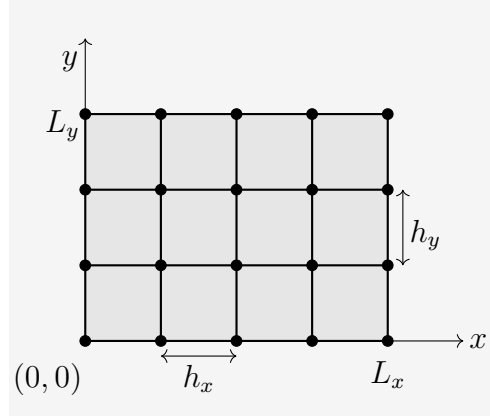
We have seen in Section 6.3 how to discretise second-order derivatives in 1D. In 2D, we then logically have for a function $f(x, y)$

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) = \frac{f(x_0 + h_x, y_0) - 2f(x_0, y_0) + f(x_0 - h_x, y_0)}{h_x^2} + \mathcal{O}(h_x^2) \tag{6.75}$$

$$\frac{\partial^2 f}{\partial y^2}(x_0, y_0) = \frac{f(x_0, y_0 + h_y) - 2f(x_0, y_0) + f(x_0, y_0 - h_y)}{h_y^2} + \mathcal{O}(h_y^2) \tag{6.76}$$

What about mixed derivatives? Since these are combinations of first-order derivatives, we can straightforwardly discretise them:

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$$

$$= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)(x_0, y_0)$$

$$= \frac{\partial}{\partial x}\left(\frac{f(x_0, y_0 + h_y) - f(x_0, y_0 - h_y)}{2h_y}\right)$$

$$= \frac{1}{2h_y}\frac{\partial f}{\partial x}(x_0, y_0 + h_y) - \frac{1}{2h_y}\frac{\partial f}{\partial x}(x_0, y_0 - h_y)$$

$$= \frac{1}{2h_y}\frac{f(x_0 + h_x, y_0 + h_y) - f(x_0 - h_x, y_0 + h_y)}{2h_x} - \frac{1}{2h_y}\frac{f(x_0 + h_x, y_0 - h_y) - f(x_0 - h_x, y_0 - h_y)}{2h_x}$$

$$= \frac{f(x_0 + h_x, y_0 + h_y) - f(x_0 - h_x, y_0 + h_y) - f(x_0 + h_x, y_0 - h_y) + f(x_0 - h_x, y_0 - h_y)}{2h_x h_y} + \mathcal{O}(h_x^2, h_y^2)$$

From 1D to 2D

INSERT TEXT

(tikz_needicon.tex)

196

1D

2D
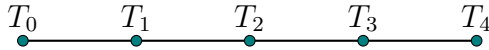
$$\vec{T} = \begin{pmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix}$$
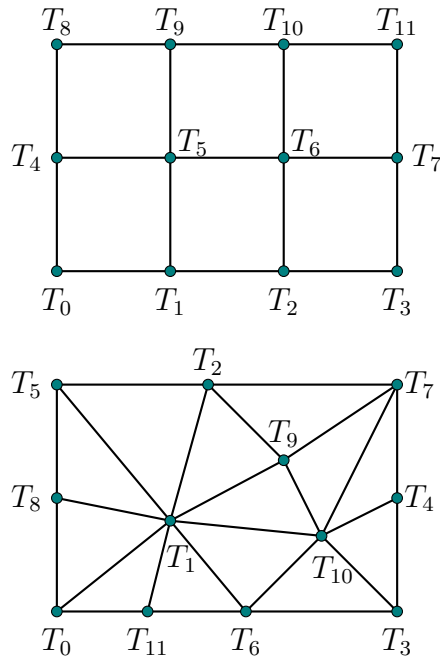
$$\vec{T} = \begin{pmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \\ T_7 \\ T_8 \\ T_9 \\ T_{10} \\ T_{11} \end{pmatrix}$$

Also, here is a rather handy code snippet which should allow you to make nice plots of the coming exercises.

```
filename = 'solution_{:04d}.pdf'.format(istep)
fig = plt.figure ()
#ax = fig.gca(projection='3d')
ax = fig.add_subplot(projection='3d')
ax.plot_surface(x.reshape ((nny,nnx)),y.reshape((nny,nnx)),T.reshape((nny,nnx)),color = 'd
ax.set_xlabel ( 'X [ m ] ')
ax.set_ylabel ( 'Y [ m ] ')
ax.set_zlabel ( ' Temperature  [ C ] ')
plt.title('Timestep  %.2d' %(istep),loc='right')
plt.grid ()
plt.savefig(filename)
#plt.show ()
plt.close()
```

197

6.8 Solving the 2d diffusion equation

We now revisit the transient heat equation, this time with sources/sinks for 2D problems. In the absence of advective heat transport, the heat equation is

$$\rho C_p \frac{\partial T}{\partial t} = \vec{\nabla} \cdot k \vec{\nabla} T + Q \tag{6.77}$$

where Q is the radiogenic heat production. It simply writes as follows when Cartesian coordinates are used:

$$\rho C_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) + Q \tag{6.78}$$

If the heat conductivity is constant in space (and so are the other coefficients), it writes:

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + \tilde{Q} \tag{6.79}$$

with $\tilde{Q} = Q/\rho C_p$. In order to solve this equation over the Cartesian domain of size $L_x \times L_y$ we need to generate a mesh as shown hereunder:
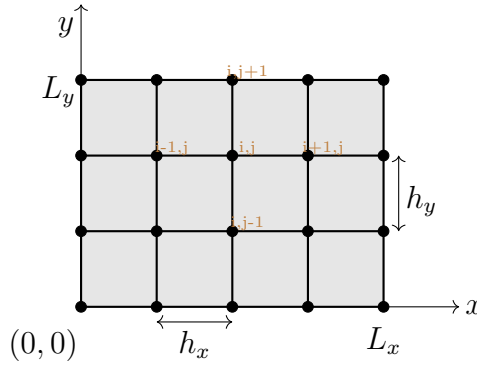
The spacing between the nodes in the x-direction is h_x and h_y is the spacing between the nodes in the y direction. There are now $nnp = nnx \times nny$ nodes in total. The above grid is characterised by $i = 0, 1, 2, 3, 4$ and $j = 0, 1, 2, 3$ and counts in total 20 nodes.

In one dimension, the subscript indicated the node i. In two dimensions we therefore need two indices i and j to identify a node, so that the temperature at node i, j at time n is denoted $T_{i,j}^n$.

One question remains: should we number nodes row by row ? column by column ? randomly ? These three approaches are shown hereunder:
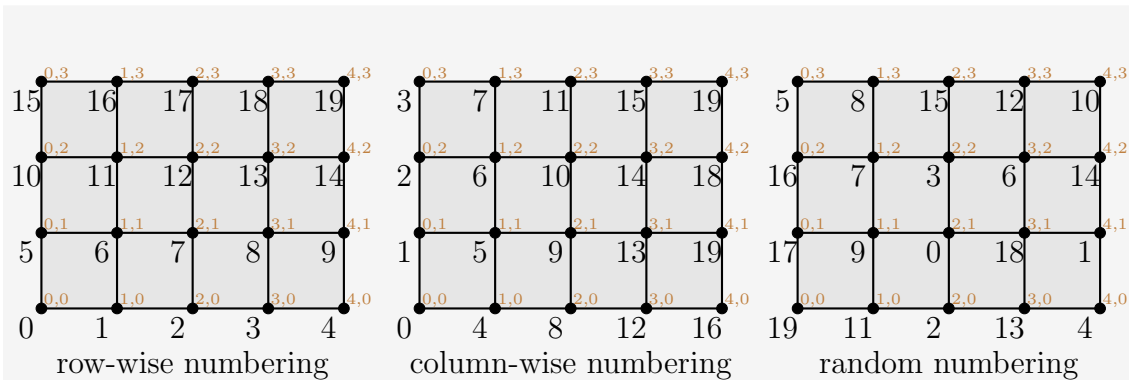
This is a critical point because the discretised PDE is formulated as a function of $T_{i,j}$ with $i = 0, \ldots nnx - 1$ and $j = 0, \ldots nny - 1$ but the vector $\vec{T}$ containing all these values (encountered in implicit methods) is indexed by a single index $k = 0, \ldots nnp - 1$. The numbering strategy determines

how easy it is to go from (i, j) to k and vice versa. Very concretely again, where should $T_{3,4}$ be placed in the global vector of unknowns $\vec{T}$?

At the same time we cannot do away with i, j indices because these are needed to locate the direct neighbours of any node and allow to form discrete derivatives.

We then need a (preferably simple/straightforward) 'function' which associates to every (i, j) a global index k. For the first grid with row-wise numbering, we have $0 \le i \le 4$, $0 \le j \le 3$ and $0 \le k \le 19$. It follows that

$$k(i, j) = j \cdot nnx + i \tag{6.80}$$

This is easy to verify: $i = 3$ and $j = 2$ indeed corresponds to node # 13, $i = 4$ and $j = 1$ corresponds to node # 9, etc ...
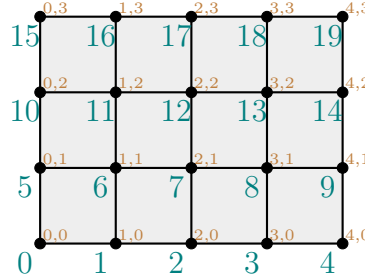
Exercise FDM-9

In a new code declare and assign values to nnx and nny. Compute nnp. Set $L_x = 7$ and $L_y = 6$. Compute h_x and h_y.
Declare two arrays $xcoords$ and $ycoords$ which will contain the x and y coordinates of all nnp nodes.
By means of two nested for loops compute these coordinates & fill both arrays.
Visualise the nodes with matplotlib.
Tip: Make sure your code works for various combinations of nnx and nny.

6.8.1 Explicit scheme

The simplest approach is an FTCS (forward time, centered space) explicit method like in 1D:

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} = \kappa \left(\frac{T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n}{h_x^2} + \frac{T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n}{h_y^2} \right) + \tilde{Q}_{i,j}^n \tag{6.81}$$

where we have assumed that the source term $\tilde{Q}$ can depend of space coordinates and therefore appears as $\tilde{Q}_{i,j}$ in the equation. We define s_x and s_y as follows:

$$s_x = \frac{\kappa \delta t}{h_x^2} \qquad s_y = \frac{\kappa \delta t}{h_y^2} \tag{6.82}$$

so that

$$T_{i,j}^{n+1} = T_{i,j}^n + s_x(T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n) + s_y(T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n) + \tilde{Q}_{i,j}^n \delta t \tag{6.83}$$

or,

$$T_{k(i,j)}^{n+1} = T_{k(i,j)}^n + s_x(T_{k(i-1,j)}^n - 2T_{k(i,j)}^n + T_{k(i+1,j)}^n) + s_y(T_{k(i,j-1)}^n - 2T_{k(i,j)}^n + T_{k(i,j+1)}^n) + \tilde{Q}_{k(i,j)}^n \delta t \tag{6.84}$$

The scheme is stable for

$$\delta t \leq \frac{\min(h_x^2, h_y^2)}{2\kappa} \tag{6.85}$$

Boundary conditions can be set the usual way: for example a constant (Dirichlet) temperature at node (i, j) is given by

$$T_{i,j} = T_{bc} \tag{6.86}$$

where T_{bc} is the prescribed temperature.

Exercise FDM-10

A simple (time-dependent) analytical solution for the temperature equation exists for the case that the initial temperature field is

$$T(x, y, t = 0) = T_0 + T_{max} \exp\left[-\frac{x^2 + y^2}{\sigma^2}\right] \tag{6.87}$$

where T_{max} is the maximum amplitude of the temperature perturbation at $(x, y) = (0, 0)$ and σ its half-width.
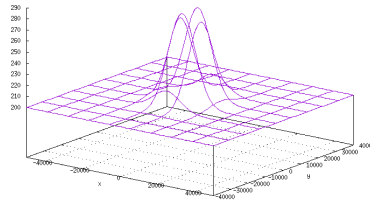
initial temperature field

The solution of the time-dependent PDE is

$$T(x, y, t) = T_0 + \frac{T_{max}}{1 + 4t\kappa/\sigma^2} \exp\left[-\frac{x^2 + y^2}{\sigma^2 + 4t\kappa}\right] \tag{6.88}$$

Set L_x=100km and $L_y = 80$km, $\kappa = 10^{-6}$, $\tilde{Q} = 0$, $T_{max} = 100°$, $T_0 = 200°$, and $\sigma = 10^4$m.
Use the previous exercise to generate a $nnx \times nny$ grid in the $[-L_x/2, L_x/2] \times [-L_y/2, L_y/2]$ domain.
Write a function which takes x, y, t, T_0, T_{max}, κ and σ as argument and returns the analytical temperature value.
Write a an explicit FDM code which solves the 2D diffusion equation. At each time step prescribe on the boundary the analytical solution.
Plot the error field (i.e. the obtained solution minus the analytical one) in the domain at a few times.

6.8.2 Implicit scheme

If we now employ a fully implicit, unconditionally stable discretization scheme, the discretised PDE becomes:

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} = \kappa \left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2} \right) + \frac{Q_{i,j}^n}{\rho C_p} \qquad (6.89)$$

Rearranging terms with $n+1$ on the left and terms with n on the right hand side gives

$$-s_x \, T_{i+1,j}^{n+1} - s_y \, T_{i,j+1}^{n+1} + (1 + 2s_x + 2s_y) \, T_{i,j}^{n+1} - s_x \, T_{i-1,j}^{n+1} - s_y \, T_{i,j-1}^{n+1} = T_{i,j}^n + \tilde{Q}_{i,j}^n \delta t \qquad (6.90)$$

or

$$-s_x \, T_{k(i+1,j)}^{n+1} - s_y \, T_{k(i,j+1)}^{n+1} + (1 + 2s_x + 2s_y) \, T_{k(i,j)}^{n+1} - s_x \, T_{k(i-1,j)}^{n+1} - s_y \, T_{k(i,j-1)}^{n+1} = T_{k(i,j)}^n + \tilde{Q}_{k(i,j)}^n \delta t \quad (6.91)$$

which here again yields a linear system of equations written $\boldsymbol{A} \cdot \vec{T} = \vec{b}$ where $\boldsymbol{A}$ is a $(nnp \times nnp)$ matrix.
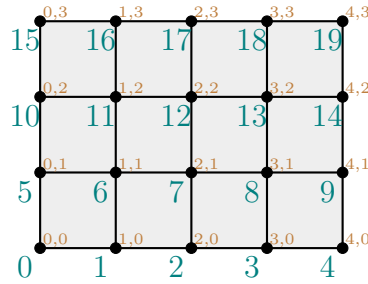
For simplicity boundary conditions are $T(x, y) = T_{bc}$ on all sides, so all nodes on the boundary have a prescribed T_{bc} temperature[10]:

$$
\begin{aligned}
T_{0,0} = T_0 &= T_{bc} \\
T_{1,0} = T_1 &= T_{bc} \\
T_{2,0} = T_2 &= T_{bc} \\
T_{3,0} = T_3 &= T_{bc} \\
T_{4,0} = T_4 &= T_{bc} \\
T_{0,1} = T_5 &= T_{bc} \\
T_{4,1} = T_9 &= T_{bc} \\
T_{0,2} = T_{10} &= T_{bc} \\
T_{4,2} = T_{14} &= T_{bc} \\
T_{0,3} = T_{15} &= T_{bc} \\
T_{1,3} = T_{16} &= T_{bc} \\
T_{2,3} = T_{17} &= T_{bc} \\
T_{3,3} = T_{18} &= T_{bc} \\
T_{4,3} = T_{19} &= T_{bc}
\end{aligned}
$$

In what follows we assume for simplicity and conciseness of notation that $h_x = h_y = h$ so that $s_x = s_y = s$. The discretised PDE equation will now be applied to the interior nodes:

- For node $k = 6$ ($i = 1, j = 1$):

$$
\begin{aligned}
-sT_{2,1}^{n+1} - sT_{1,2}^{n+1} + (1+4s)T_{1,1}^{n+1} - sT_{0,1}^{n+1} - sT_{1,0}^{n+1} &= T_{1,1}^n + \tilde{Q}_{1,1}^n \delta t \\
\Rightarrow \quad -sT_7^{n+1} - sT_{11}^{n+1} + (1+4s)T_6^{n+1} - sT_5^{n+1} - sT_1^{n+1} &= T_6^n + \tilde{Q}_6^n \delta t \qquad (6.92)
\end{aligned}
$$

---

[10]We assume here again that these boundary conditions do not change with time.

- For node $k = 7$ $(i = 2, j = 1)$:

$$-sT_{3,1}^{n+1} - sT_{2,2}^{n+1} + (1+4s)T_{2,1}^{n+1} - sT_{1,1}^{n+1} - sT_{2,0}^{n+1} = T_{2,1}^n + \tilde{Q}_{2,1}^n \delta t$$
$$\Rightarrow \quad -sT_8^{n+1} - sT_{12}^{n+1} + (1+4s)T_7^{n+1} - sT_6^{n+1} - sT_2^{n+1} = T_7^n + \tilde{Q}_7^n \delta t \qquad (6.93)$$

- For node $k = 8$ $(i = 3, j = 1)$:

$$-sT_{4,1}^{n+1} - sT_{3,2}^{n+1} + (1+4s)T_{3,1}^{n+1} - sT_{2,1}^{n+1} - sT_{3,0}^{n+1} = T_{3,1}^n + \tilde{Q}_{3,1}^n \delta t$$
$$\Rightarrow \quad -sT_9^{n+1} - sT_{13}^{n+1} + (1+4s)T_8^{n+1} - sT_7^{n+1} - sT_3^{n+1} = T_8^n + \tilde{Q}_8^n \delta t \qquad (6.94)$$

- For node $k = 11$ $(i = 1, j = 2)$:

$$-sT_{2,2}^{n+1} - sT_{1,3}^{n+1} + (1+4s)T_{1,2}^{n+1} - sT_{0,2}^{n+1} - sT_{2,1}^{n+1} = T_{1,2}^n + \tilde{Q}_{1,2}^n \delta t$$
$$\Rightarrow \quad -sT_{12}^{n+1} - sT_{16}^{n+1} + (1+4s)T_{11}^{n+1} - sT_{10}^{n+1} - sT_6^{n+1} = T_{11}^n + \tilde{Q}_{11}^n \delta t \qquad (6.95)$$

- For node $k = 12$ $(i = 2, j = 2)$:

$$-sT_{3,2}^{n+1} - sT_{2,3}^{n+1} + (1+4s)T_{2,2}^{n+1} - sT_{1,2}^{n+1} - sT_{2,1}^{n+1} = T_{2,2}^n + \tilde{Q}_{2,2}^n \delta t$$
$$\Rightarrow \quad -sT_{13}^{n+1} - sT_{17}^{n+1} + (1+4s)T_{12}^{n+1} - sT_{11}^{n+1} - sT_7^{n+1} = T_{12}^n + \tilde{Q}_{12}^n \delta t \qquad (6.96)$$

- For node $k = 13$ $(i = 3, j = 2)$:

$$-sT_{4,2}^{n+1} - sT_{3,3}^{n+1} + (1+4s)T_{3,2}^{n+1} - sT_{2,2}^{n+1} - sT_{3,1}^{n+1} = T_{3,2}^n + \tilde{Q}_{3,2}^n \delta t$$
$$\Rightarrow \quad -sT_{14}^{n+1} - sT_{18}^{n+1} + (1+4s)T_{13}^{n+1} - sT_{12}^{n+1} - sT_8^{n+1} = T_{13}^n + \tilde{Q}_{13}^n \delta t \qquad (6.97)$$

Putting it all together yields the following linear system:

$$
\underbrace{\begin{pmatrix}
1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . & . & . & . & . & . & . & . \\
. & . & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . & . & . & . & . & . & . \\
. & . & . & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . & . & . \\
. & . & . & . & . & . & . & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . & . \\
. & . & . & . & . & . & . & . & -s & . & . & . & -s & 1+4s & -s & . & . & . & -s & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1
\end{pmatrix}}_{\boldsymbol{A}}
\cdot
\underbrace{\begin{pmatrix}
T_0^{n+1} \\ T_1^{n+1} \\ T_2^{n+1} \\ T_3^{n+1} \\ T_4^{n+1} \\ T_5^{n+1} \\ T_6^{n+1} \\ T_7^{n+1} \\ T_8^{n+1} \\ T_9^{n+1} \\ T_{10}^{n+1} \\ T_{11}^{n+1} \\ T_{12}^{n+1} \\ T_{13}^{n+1} \\ T_{14}^{n+1} \\ T_{15}^{n+1} \\ T_{16}^{n+1} \\ T_{17}^{n+1} \\ T_{18}^{n+1} \\ T_{19}^{n+1}
\end{pmatrix}}_{\vec{T}}
=
\underbrace{\begin{pmatrix}
T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_6^n + \tilde{Q}_6 \delta t \\ T_7^n + \tilde{Q}_7 \delta t \\ T_8^n + \tilde{Q}_8 \delta t \\ T_{bc} \\ T_{bc} \\ T_{11}^n + \tilde{Q}_{11} \delta t \\ T_{12}^n + \tilde{Q}_{12} \delta t \\ T_{13}^n + \tilde{Q}_{13} \delta t \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc} \\ T_{bc}
\end{pmatrix}}_{\vec{b}}
$$

Note that we now have five 'diagonals' filled with non-zero entries as opposed to three diagonals in the 1D case.

Note again that this is a simplified matrix since we assumed that $s_x = s_y$ (see expressions at the beginning of this section).

Let us now focus on a special case: we wish to solve the steady state diffusion equation. The approach is then very similar, but we must discard the time derivative term so that the discretised PDE becomes:

$$0 = \kappa \left(\frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{h_x^2} + \frac{T_{i,j-1} - 2T_{i,j} + T_{i,j+1}}{h_y^2} \right) + \frac{Q_{i,j}}{\rho C_p} \qquad (6.98)$$

Note that we have removed the superscript n and $n+1$ since there is no time discretisation.

Rearranging terms, and assuming again that $h_x = h_y$ for simplicity[11]:

$$-T_{i+1,j} - T_{i,j+1} + 4\,T_{i,j} - T_{i-1,j} - T_{i,j-1} = \frac{\tilde{Q}_{i,j} h^2}{\kappa} \qquad (6.99)$$

We would then follow the regular procedure of writing this relationship at nodes that are not on the boundary. In the end we would obtain a matrix with an identical structure as before, but with slightly different terms (esp. on the diagonal).

A final remark: all this is valid for a a constant heat conductivity (which allowed us to take it out of the gradient terms in the original PDE and form the heat diffusivity term κ - also constant in space). If not the expression presented in Eq. (6.34) must be used.

Exercise FDM-11

Same exercise as exercise FDM-10, but now with implicit method.

Looking at this matrix, it is clear that this approach is sub-optimal: for such a small grid counting 20 nodes, the boundary conditions enforce the temperature on 14 of them, so that these temperatures should/could be removed from the list of unknowns, leaving a vector of unknowns $\vec{T}$ of size 6 (the number of nodes which are not on the boundary). As a consequence, we would have to solve a 6×6 linear system, as opposed to a 20×20 one!

In this case, we focus again on nodes 6,7,8,11,12,13. we start from

$$-sT_7^{n+1} - sT_{11}^{n+1} + (1+4s)T_6^{n+1} - sT_5^{n+1} - sT_1^{n+1} = T_6^n + \tilde{Q}_6^n \delta t \qquad (6.100)$$

but we know that the boundary conditions impose that $T_1 = 0$ and $T_5 = 0$ so that the equation above simplifies to:

$$-sT_7^{n+1} - sT_{11}^{n+1} + (1+4s)T_6^{n+1} = T_6^n + \tilde{Q}_6^n \delta t \qquad (6.101)$$

These 6 equations can finally be combined in the expected smaller linear system:

$$\underbrace{\begin{pmatrix} 1+4s & -s & . & -s & . & . \\ -s & 1+4s & -s & . & -s & . \\ . & -s & 1+4s & . & . & -s \\ -s & . & -s & 1+4s & -s & . \\ . & -s & . & -s & 1+4s & -s \\ . & . & -s & . & -s & 1+4s \end{pmatrix}}_{\boldsymbol{A}} \cdot \underbrace{\begin{pmatrix} T_6^{n+1} \\ T_7^{n+1} \\ T_8^{n+1} \\ T_{11}^{n+1} \\ T_{12}^{n+1} \\ T_{13}^{n+1} \end{pmatrix}}_{\vec{T}} = \underbrace{\begin{pmatrix} T_6^n + \tilde{Q}_6^n \delta t \\ T_7^n + \tilde{Q}_7^n \delta t \\ T_8^n + \tilde{Q}_8^n \delta t \\ T_{11}^n + \tilde{Q}_{11}^n \delta t \\ T_{12}^n + \tilde{Q}_{12}^n \delta t \\ T_{13}^n + \tilde{Q}_{13}^n \delta t \end{pmatrix}}_{\vec{b}} \qquad (6.102)$$

---

[11]We no longer use the s_x and s_y parameters since those contain the term δt which is here meaningless

Note that is the boundary values had not been zero they would have found their way to the right hand side vector.

The Crank-Nicolson version of the implicit scheme is then as follows:

$$
\begin{aligned}
\frac{T_{i,j}^{n+1} - T_{i,j}{}^{n}}{\delta t} \;&=\; \frac{1}{2}\kappa \left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2} \right) \\
&+\; \frac{1}{2}\kappa \left(\frac{T_{i-1,j}^{n} - 2T_{i,j}^{n} + T_{i+1,j}^{n}}{h_x^2} + \frac{T_{i,j-1}^{n} - 2T_{i,j}^{n} + T_{i,j+1}^{n}}{h_y^2} \right)
\end{aligned}
\tag{6.103}
$$

The implementation of this method will require from you to bring all the terms in T^{n+1} to the left of the equal sign while all the terms in T^n are assumed to be known and therefore find their way into the right hand side.

Likewise, the Lax-Friedrichs method is as follows:

$$
\frac{T_{i,j}^{n+1} - \frac{1}{4}\left(T_{i-1,j}^{n} + T_{i+1,j}^{n} + T_{i,j-1}^{n} + T_{i,j+1}^{n} \right)}{\delta t} = \kappa \left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2} \right) + \frac{Q_{i,j}^{n}}{\rho C_p}
\tag{6.104}
$$

Rearranging terms with $n+1$ on the left and terms with n on the right hand side gives

6.8.3 The 9-point stencil for the Laplace operator

What follows is mostly borrowed from Wikipedia[12].

If we discretize the 2D Laplacian by using central-difference methods, we obtain the commonly used five-point stencil, represented by the following convolution kernel:

$$
D = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}
$$

or,

$$
DT_{i,j} = \frac{1}{h^2}(T_{i-1,j} + T_{i+1,j}, + T_{i,j-1} + T_{i,j+1}, - 4T_{i,j})
$$

Even though it is simple to obtain and computationally lighter, the central difference kernel possess an undesired intrinsic anisotropic property, since it doesn't take into account the diagonal neighbours.

The two most commonly used isotropic nine-point stencils are displayed below, in their convolution kernel forms. They can be obtained by the following formula

$$
D = (1-\gamma)\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \gamma \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & -2 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}
$$

The first one is known by Oono-Puri,and it is obtained when $\gamma = 1/2$:

$$
D = \begin{bmatrix} 1/4 & 2/4 & 1/4 \\ 2/4 & -12/4 & 2/4 \\ 1/4 & 2/4 & 1/4 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 2 & -12 & 2 \\ 1 & 2 & 1 \end{bmatrix}
$$

The second one is known by Patra-Karttunen or Mehrstellen, and it is obtained when $\gamma = 1/3$:

$$
D = \begin{bmatrix} 1/6 & 4/6 & 1/6 \\ 4/6 & -20/6 & 4/6 \\ 1/6 & 4/6 & 1/6 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix}
$$

---

[12]https://en.wikipedia.org/wiki/Nine-point_stencil

or,

$$\vec{\nabla}^2 T_{i,j} = \frac{1}{6h^2}(T_{i+1,j+1} + T_{i-1,j+1} + T_{i+1,j-1}, + T_{i-1,j-1} + 4(T_{i+1,j} + T_{i-1,j}) + 4(T_{i,j+1} + T_{i,j-1}) - 20T_{i,j})$$

Both are isotropic forms of discrete Laplacian, and in the limit of small h, they all become equivalent. This form is the one we find in LeVeque [824, p64].

6.8.4   The alternating-direction-implicit (ADI) technique

This is borroed from Anderson's book, section 6.7.

We start again from the Crank-Nicolson version of the implicit scheme:

$$
\begin{aligned}
\frac{T_{i,j}^{n+1} - T_{i,j}{}^n}{\delta t} \quad &= \quad \frac{1}{2}\kappa \left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2} \right) \\
&+ \quad \frac{1}{2}\kappa \left(\frac{T_{i-1,j}^{n} - 2T_{i,j}^{n} + T_{i+1,j}^{n}}{h_x^2} + \frac{T_{i,j-1}^{n} - 2T_{i,j}^{n} + T_{i,j+1}^{n}}{h_y^2} \right) \quad (6.105)
\end{aligned}
$$

Although it is usable 'as is', there was an incentive a few decennia to keep the matrix tridiagonal (as opposed to pentadiagonal in this case) so that the Thomas algorithm[13] could be used.

The ADI method arrives at the solution in a two-step process, where intermediate values of T are found at an intermediate time $t + \delta t/2$, as follows. For each step a tridiagonal matrix is generated.

In the first step replace the spatial derivative with central differences where only the x derivative is treated implicitely:

$$\frac{T_{i,j}^{n+1/2} - T_{i,j}{}^n}{\delta t/2} \quad = \quad \kappa \frac{T_{i-1,j}^{n+1/2} - 2T_{i,j}^{n+1/2} + T_{i+1,j}^{n+1/2}}{h_x^2} + \kappa \frac{T_{i,j-1}^{n} - 2T_{i,j}^{n} + T_{i,j+1}^{n}}{h_y^2} \quad (6.106)$$

The second step of the ADI scheme takes the solution to $t + \delta t$, using the known values at time $t + \delta t/2$:

$$\frac{T_{i,j}^{n+1} - T_{i,j}{}^{n+1/2}}{\delta t/2} \quad = \quad \kappa \frac{T_{i-1,j}^{n+1/2} - 2T_{i,j}^{n+1/2} + T_{i+1,j}^{n+1/2}}{h_x^2} + \kappa \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2} \quad (6.107)$$

The ADI method is second order in time and space.

6.9   Solving the 2d advection-diffusion equation

So far, we have mainly focused on the diffusion equation in a non-moving flow (relevant for the case of a dike intrusion cooling off or for a lithosphere which remains undeformed).

We now want to consider problems where material moves during the time period under consideration and takes temperature anomalies with it (e.g. a plume rising through a convecting mantle). If the numerical grid remains fixed in the background, the hot temperatures should be moved to different grid points at each time step.

We start again from the heat transport equation of Section 3.6:

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot k \vec{\nabla} T + Q \quad (6.108)$$

---

[13]https://en.wikipedia.org/wiki/Tridiagonal_matrix_algorithm. See also appendix A of Anderson's book.

We have previously dealt with the one-dimensional Cartesian coordinates equation:

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + Q \tag{6.109}$$

and we now turn to the two-dimensional equation:

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} \right) = \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left(k \frac{\partial T}{\partial y} \right) + Q \tag{6.110}$$

As before, asumming that k is constant in space we can rewrite the equation as a function of the heat diffusivity κ:

$$\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) + Q \tag{6.111}$$

Since we have already seen how to deal with 'pure' diffusion equations in the previous section, let us now turn to the 'pure' advection equations:

$$\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = 0 \tag{6.112}$$

where we assume the velocity field $\vec{v} = (u, v)$ known.

Even though the equations appear simple, it is quite tricky to solve them accurately, more so than for the diffusion problem. This is particularly the case if there are large gradients in the quantity that is to be advected.

We have seen how to deal with the time derivative (explicit, implicit) and with the first order space derivative (forward, backward or central). Let us consider again the FTCS scheme (Forward in Time, Central in Space).

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + u_{i,j} \frac{T_{i+1,j}^n - T_{i-1,j}^n}{2h_x} + v_{i,j} \frac{T_{i,j+1}^n - T_{i,j-1}^n}{2h_y} = 0 \tag{6.113}$$

The fully implicit version is then as follows:

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + u_{i,j} \frac{T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}}{2h_x} + v_{i,j} \frac{T_{i,j+1}^{n+1} - T_{i,j-1}^{n+1}}{2h_y} = 0 \tag{6.114}$$

or,

$$T_{i,j}^{n+1} + \frac{u_{i,j}\delta t}{2h_x}(T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}) + \frac{v_{i,j}\delta t}{2h_y}(T_{i,j+1}^{n+1} - T_{i,j-1}^{n+1}) = T_{i,j}^n$$

The terms on the left will form five diagonals in the matrix while the term on the right is the right hand side.

The Crank-Nicolson approach is then easily derived by taking:

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + \frac{u_{i,j}}{2} \frac{T_{i+1,j}^n - T_{i-1,j}^n}{2h_x} + \frac{v_{i,j}}{2} \frac{T_{i,j+1}^n - T_{i,j-1}^n}{2h_y} + \frac{u_{i,j}}{2} \frac{T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}}{2h_x} + \frac{v_{i,j}}{2} \frac{T_{i,j+1}^{n+1} - T_{i,j-1}^{n+1}}{2h_y} = 0 \tag{6.115}$$

Let us now put advection and diffusion together: Let us consider again the FTCS scheme (Forward in Time, Central in Space).

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + u_{i,j} \frac{T_{i+1,j}^n - T_{i-1,j}^n}{2h_x} + v_{i,j} \frac{T_{i,j+1}^n - T_{i,j-1}^n}{2h_y} = \kappa \left(\frac{T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n}{h_x^2} + \frac{T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n}{h_y^2} \right) \tag{6.116}$$

We multiply by δt:

$$T_{i,j}^{n+1} - T_{i,j}^n + \underbrace{\frac{u_{i,j}\delta t}{2h_x}}_{c_{i,j}^u}(T_{i+1,j}^n - T_{i-1,j}^n) + \underbrace{\frac{v_{i,j}\delta t}{2h_y}}_{c_{i,j}^v}(T_{i,j+1}^n - T_{i,j-1}^n)$$

$$= \underbrace{\frac{\kappa\delta t}{h_x^2}}_{s_x}\left(T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n\right) + \underbrace{\frac{\kappa\delta t}{h_y^2}}_{s_y}\left(T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n\right) \qquad (6.117)$$

and in the end:

$$T_{i,j}^{n+1} - T_{i,j}^n + c_{i,j}^u(T_{i+1,j}^n - T_{i-1,j}^n) + c_{i,j}^v(T_{i,j+1}^n - T_{i,j-1}^n)$$
$$= s_x\left(T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n\right) + s_y\left(T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n\right) \qquad (6.118)$$

or,

$$T_{i,j}^{n+1} = T_{i,j}^n - c_{i,j}^u(T_{i+1,j}^n - T_{i-1,j}^n) - c_{i,j}^v(T_{i,j+1}^n - T_{i,j-1}^n)$$
$$+ s_x\left(T_{i-1,j}^n - 2T_{i,j}^n + T_{i+1,j}^n\right) + s_y\left(T_{i,j-1}^n - 2T_{i,j}^n + T_{i,j+1}^n\right) \qquad (6.119)$$

and in the end the explicit form of the 2d advection-diffusion equation is:

$$T_{i,j}^{n+1} = (1-2s_x-2s_y)T_{i,j}^n + (-c_{i,j}^u+s_x)T_{i+1,j}^n + (c_{i,j}^u+s_x)T_{i-1,j}^n + (-c_{i,j}^v+s_y)T_{i,j+1}^n + (c_{i,j}^v+s_y)T_{i,j-1}^n$$

Likewise the implicit form of the 2d advection-diffusion equation is:

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + u_{i,j}\frac{T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}}{2h_x} + v_{i,j}\frac{T_{i,j+1}^{n+1} - T_{i,j-1}^{n+1}}{2h_y} = \kappa\left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j-1}^{n+1} - 2T_{i,j}^{n+1} + T_{i,j+1}^{n+1}}{h_y^2}\right)$$

which ultimately leads to write

$$(1+2s_x+2s_y)T_{i,j}^{n+1} + (c_{i,j}^u-s_x)T_{i+1,j}^{n+1} + (-c_{i,j}^u-s_x)T_{i-1,j}^{n+1} + (c_{i,j}^v-s_y)T_{i,j+1}^{n+1} + (-c_{i,j}^v-s_y)T_{i,j-1}^{n+1} = T_{i,j}^n$$

or

$$(1+2s_x+2s_y)T_k^{n+1} + (c_k^u-s_x)T_{kright}^{n+1} + (-c_k^u-s_x)T_{kleft}^{n+1} + (c_k^v-s_y)T_{ktop}^{n+1} + (-c_k^v-s_y)T_{kbot}^{n+1} = T_k^n$$

Note that if one solves the steady state equation then this becomes

$$(2s_x+2s_y)T_k + (c_k^u-s_x)T_{kright} + (-c_k^u-s_x)T_{kleft} + (c_k^v-s_y)T_{ktop} + (-c_k^v-s_y)T_{kbot} = 0$$

If Dirichlet boundary conditions are prescribed on all sides then the equation above can be used on the interior nodes and once the matrix and the rhs are built the linear system can be solved.

If however a boundary (or a portion of the boundary) is not set to a given value (Dirichlet boundary condition) but instead is free of any constraint, then one must be careful with the centered scheme. Instead backward or forward schemes might be needed.

Let us assume that no boundary condition is applied at the bottom of the domain (characterised by $j=0$). The centered approach for the $\partial^2 T/\partial x^2$ is fine (but not in the corners!) but one must use a forward scheme for the y derivative:

$$\frac{\partial^2 T}{\partial y^2} \simeq \frac{T_{i,j} - 2T_{i,j+1} + T_{i,j+2}}{h_y^2}$$

and of course also for the advection term:

$$\frac{\partial T}{\partial y} \simeq \frac{T_{i,j+1} - T_{i,j}}{h_y}$$

In this case the discretised advection-diffusion equation will be (in the implicit case)

$$\frac{T_{i,j}^{n+1} - T_{i,j}^n}{\delta t} + u_{i,j}\frac{T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}}{2h_x} + v_{i,j}\frac{T_{i,j+1}^{n+1} - T_{i,j}^{n+1}}{h_y} = \kappa\left(\frac{T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}}{h_x^2} + \frac{T_{i,j}^{n+1} - 2T_{i,j+1}^{n+1} + T_{i,j+2}^{n+1}}{h_y^2}\right)$$

We multiply by δt:

$$T_{i,j}^{n+1} - T_{i,j}^n + \underbrace{\frac{u_{i,j}\delta t}{2h_x}}_{c_{i,j}^u}(T_{i+1,j}^{n+1} - T_{i-1,j}^{n+1}) + 2\underbrace{\frac{v_{i,j}\delta t}{2h_y}}_{c_{i,j}^v}(T_{i,j+1}^{n+1} - T_{i,j}^{n+1})$$

$$= \underbrace{\frac{\kappa\delta t}{h_x^2}}_{s_x}\left(T_{i-1,j}^{n+1} - 2T_{i,j}^{n+1} + T_{i+1,j}^{n+1}\right) + \underbrace{\frac{\kappa\delta t}{h_y^2}}_{s_y}\left(T_{i,j}^{n+1} - 2T_{i,j+1}^{n+1} + T_{i,j+2}^{n+1}\right) \quad (6.120)$$

and ultimately obtain

$$(1 - 2c_{i,j}^v + 2s_x - s_y)T_{i,j}^{n+1} + (c_{i,j}^u - s_x)T_{i+1,j}^{n+1} + (-c_{i,j}^u - s_x)T_{i-1,j}^{n+1} + (2c_{i,j}^v + 2s_y)T_{i,j+1}^{n+1} + (-s_y)T_{i,j+2}^{n+1} = T_{i,j}^n$$

The same reasoning holds for any boundary, one then must be careful about whether backward or forward approximations are needed for dervatives with respect to x or y.

<div align="center">

Exercise FDM-11

</div>

We wish to compute the advection of a product-cosine hill in a prescribed velocity field. The initial temperature is:

$$T_0(x,y) = \begin{cases} \frac{1}{4}\left(1 + \cos\pi\frac{x-x_c}{\sigma}\right)\left(1 + \cos\pi\frac{y-y_c}{\sigma}\right) & \text{if } (x-x_c)^2 + (y-y_c)^2 \leq \sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

The boundary conditions are $T(x,y) = 0$ on all four sides of the unit square domain. In what follows we set $x_c = y_c = 2/3$ and $\sigma = 0.2$. The velocity field is analytically prescribed: $\vec{v} = (-(y - L_y/2), +(x - L_x/2))$. Resolution is set to 31×31 nodes.

The timestep is set to $\delta t = 2\pi/200$ and we wish to carry out 200 timesteps so that the cone does a 2π rotation.

See Stone 43 for results/figures of this experiment obtained with Finite Elements.

Implement this with the FTCS method. What do you observe? What happens when you decrease the value of δt?

Bonus: Lax method, Crank-Nicolson method.

Exercise FDM-12

NOT FOR 2024!

Redo exercise FDM-6 in a unit square domain. The temperature field at $t = 0$ is given by $T(x, y) = 1$ for $x < 0.25$ and $T(x, y) = 0$ otherwise. The prescribed velocity is $\vec{v} = (1, 0)$ and we set $nnx = nny = 51$. Boundary conditions are $T = 1$ at $x = 0$ and $T = 0$ at $x = 1$.

Program the above FTCS method. Run the model for 250 time steps with $\delta t = 0.002$. Compare the 2D solution with the previously obtained 1D solution of exercise FDM-6.

Make sure the code works in the y-direction too by rotating the initial temperature by 90° anti-clockwise, set $\vec{v} = (0, 1)$ and change boundary conditions accordingly.

Bonus: Lax method, Crank-Nicolson method.

Note to self: - CFL missing still - stensils next to boundaries missing too - add many more visual aids

6.10 FEM vs FDM?

Let us start with the 1D steady advection-diffusion equation:

$$\rho C_p u \frac{dT}{dx} - k\frac{d^2T}{dx^2} = f \qquad \text{in} \quad [0, L_x] \tag{6.121}$$

with the boundary conditions $T(x = 0) = 0$ and $T(x = L_x) = 0$.

We have seen before (see Section **??**) that the elemental matrix $\boldsymbol{K}_a$ for the advection and the elemental matrix $\boldsymbol{K}_d$ for the diffusion terms are

$$\boldsymbol{K}_a^e = \frac{\rho C_p u}{2} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \qquad \boldsymbol{K}_d^e = \frac{k}{h_x} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

where h_x is the distance between nodes in the x-direction and e denotes the element number.

Assuming that we have 5 elements (i.e. 6 nodes), the assembled 6×6 advection and diffusion matrices (before boundary conditions are applied) are:

$$\boldsymbol{K}_a = \frac{\rho C_p u}{2} \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \qquad \boldsymbol{K}_d = \frac{k}{h_x} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

The rhs is zero, so that we would have to solve $(\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T} = 0$, or:

$$\left[\frac{\rho C_p u}{2} \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} + \frac{k}{h_x} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \right] \cdot \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{pmatrix} = \vec{0}$$

Note that boundary conditions are not applied yet. Therefore the algebraic equation for an interior node i is

$$\rho C_p u \frac{T_{i+1} - T_{i-1}}{2} + \frac{k}{h_x}(-T_{i-1} + 2T_i - T_{i+1}) = 0$$

or,

$$\boxed{\rho C_p u \frac{T_{i+1} - T_{i-1}}{2h_x} - \frac{k}{h_x^2}(T_{i-1} - 2T_i + T_{i+1}) = 0} \tag{6.122}$$

However, we have seen in Section 6.3 that the second order accurate central differencing based approximate first and second derivatives written for an interior node i of a finite difference mesh with a constant node spacing of h is

$$\frac{dT}{dx}\bigg|_i \simeq \frac{T_{i+1} - T_{i-1}}{2h_x} \qquad \frac{d^2T}{dx^2} \simeq \frac{T_{i+1} - 2T_i + T_{i-1}}{h_x^2}$$

Using these approximations, the discretised formualtion of Eq. (6.121) is exactly the same as Eq. (6.122). This simple example proves that the FEM and the FDM share similarities!

It is also useful to introduce the elemental Peclet number

$$Pe = \frac{uh}{2\kappa} = \frac{uh\rho C_p}{2k}$$

211

and Eq. (6.121) becomes:

$$\frac{u}{2h_x}\left[\left(1-\frac{1}{Pe}\right)T_{i+1}+\frac{2}{Pe}T_i-\left(1+\frac{1}{Pe}\right)T_{i-1}\right]=f$$

CHECK!!!

6.11 GEO4-1427: vorticity-streamfunction project

We consider a 2D Cartesian domain. We wish to solve the incompressible isothermal Stokes equations in this domain using the vorticity-streamfunction approach (see Section 18.6). We assume that the fluids in the domain all have the same viscosity η_0 and that the gravity is vertical and pointing downwards, i.e. $\vec{g} = (g_x, g_y) = (0, -g_0)$. The density in the domain is given by a function $\rho(x, y)$ specified here after. In what follows the boundary conditions are assumed to be free slip.

- Experiment 1: the domain is a square of size 600 km. It contains two fluids, the 'mantle' and the 'slab'. The mantle has a density of 3200 kg m^{-3} and fills the domain except for a circle (the slab) of radius 100 km centered at $x = L_x/2$ and $y = \frac{2}{3}L_y$ of density 3300 kg m^{-3}. $g_0 = 10 \text{ m s}^{-2}$ and $\eta_0 = 10^{21}$ Pa s.
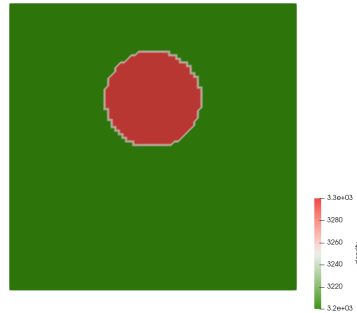
- Experiment 2: gravity field ($g_x = 0$, $g_y = 10 \text{ m s}^{-2}$) for a density structure with two vertical layers (3200 kg m^{-3} and 3300 kg m^{-3} for the left and right layers, respectively). The model size is 1500 km × 1000 km. Use a constant viscosity $\eta = 10^{21}$ Pa s for the entire model.
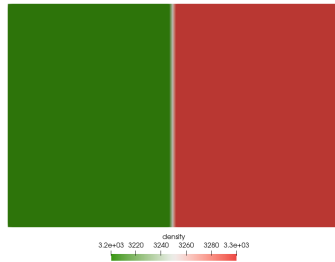
Background theory

We start from

$$\vec{\nabla}^4 \psi = \frac{1}{\eta_0} \left(-\frac{\partial(\rho g_x)}{\partial y} + \frac{\partial(\rho g_y)}{\partial x} \right) \tag{6.123}$$

where

$$\vec{\nabla}^4 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

Eq. (6.123) can be rewritten as two Poisson equations, by defining the vorticity as $\omega = \vec{\nabla}^2 \psi$:

We obtain the Poisson equation for the vorticity ω:

$$\vec{\nabla}^2 \omega = -\frac{1}{\eta_0} \left(-\frac{\partial(\rho g_x)}{\partial y} + \frac{\partial(\rho g_y)}{\partial x} \right) \tag{6.124}$$

In what follows we assume that the domain is a Cartesian box aligned with the x, y axis. The gravity is constant and vertical so that $g_x = 0$ and then we must solve

$$\vec{\nabla}^2 \omega = -\frac{g_0}{\eta_0}\frac{\partial \rho}{\partial x} \tag{6.125}$$

$$\vec{\nabla}^2 \psi = \omega \tag{6.126}$$

1. The domain under consideration is a rectangle of size $L_x \times L_y$. Create a mesh of $nnx \times nny$ nodes. Store the x and y coordinates in 2 arrays `xcoords` and `ycoords`

2. Prescribe the density ρ on the nodes of the mesh, store the values in the `rho` array. Visualise this field and produce a figure called *rho.png*.

3. Compute $\partial \rho / \partial x$ on the nodes. Visualise this field and produce a figure called *drhodx.png*. Use a centered approach if a node is inside the domain and a backward or forward approach if it is on a boundary.

4. Solve Eq. (6.125) using the FD method. Solve the linear system with a scipy solver. Boundary conditions for ω are $\omega = 0$ on all four sides of the domain. Visualise this field and produce a figure called *omega.png*

5. Solve Eq. (6.126) in a similar way, also setting $\psi = 0$ on all four sides. Visualise this field and produce a figure called *psi.png*

6. Compute the velocity components u and v at each node using

$$u = \frac{\partial \psi}{\partial y} \qquad \text{and} \qquad v = -\frac{\partial \psi}{\partial x}$$

   As in question 2, use the appropriate stencil whether a node is on a boundary or not. Test your implementation of the gradients by computing u, v at each node having set $\psi = 1$ on all nodes (test 1) and set $\psi = xy$ on all nodes (test 2). In both cases, what values of u, v do we expect? Do you recover these values? (if not there is a problem here and you should not proceed any further

   When this is correct, visualise the velocity components (in cm/year!) for the sinking sphere: produce then two files, *u.png* and *v.png*, and also a third file *vel.png* with the density overlain by velocity vectors.

7. Compute the strain rate tensor components

$$\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} \qquad \dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} \qquad \dot{\varepsilon}_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)$$

   Visualise these fields and produce then three files *exx.png*, *eyy.png* and *exy.png*.

8. Compute the effective strain rate

$$\varepsilon_e = \sqrt{\frac{1}{2}(\dot{\varepsilon}_{xx}^2 + \dot{\varepsilon}_{yy}^2) + \dot{\varepsilon}_{xy}^2}$$

   and produce a figure *ee.png*

9. Bonus: try the 9 point Laplace stencil, see Section 6.8.3.

6.12 Solving the 2d Stokes equations

6.12.1 A quick recap of the incompressible 2D Stokes equations

In the case of an incompressible and isoviscous flow, we can start from the momentum equation

$$-\vec{\nabla}p + \vec{\nabla}\cdot\boldsymbol{\tau} + \rho\vec{g} = \vec{0} \tag{6.127}$$

with[14]

$$\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}}) \tag{6.128}$$

In 2D the strain rate tensor components are given by

$$\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} \qquad = \dot{\varepsilon}_{yx} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \qquad \dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y}$$

In the end we obtain the following set of coupled PDEs for the three unknowns u, v, p:

$$-\frac{\partial p}{\partial x} + 2\frac{\partial}{\partial x}\left(\eta\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\eta(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})\right) + \rho g_x = 0 \tag{6.129}$$

$$-\frac{\partial p}{\partial y} + 2\frac{\partial}{\partial y}\left(\eta\frac{\partial v}{\partial y}\right) + \frac{\partial}{\partial x}\left(\eta(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})\right) + \rho g_y = 0 \tag{6.130}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{6.131}$$

As a side note, Becker and Kaus [71] add the term p/λ to the rhs of the incompressibility equation:

> This is a "trick" called the penalty method, which ensures that the system of equations does not become ill-posed. For this to work, λ should be sufficiently large ($\sim 10^4$ or so), so that the condition of incompressibility is approximately satisfied.

This is of course very similar to the penalty method described in Section 10.7.

Then, *if the viscosity is constant* the x-component of the momentum conservation equation

$$-\frac{\partial p}{\partial x} + 2\eta\frac{\partial\dot{\varepsilon}_{xx}}{\partial x} + 2\eta\frac{\partial\dot{\varepsilon}_{xy}}{\partial y} + \rho g_x = 0$$

becomes

$$-\frac{\partial p}{\partial x} + 2\eta\frac{\partial^2 u}{\partial x^2} + \eta\left(\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 v}{\partial x\partial y}\right)\rho g_x = 0$$

Using $\vec{\nabla}\cdot\vec{v} = 0$ to obtain $\frac{\partial u}{\partial x} = -\frac{\partial v}{\partial y}$ we arrive at

$$-\frac{\partial p}{\partial x} + \eta\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \rho g_x = 0$$

The same approach can be carried out for the y-component of the momentum conservation equation and in the end these are the three coupled equations we need to solve

$$-\frac{\partial p}{\partial x} + \eta\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \rho g_x = 0 \tag{6.132}$$

$$-\frac{\partial p}{\partial y} + \eta\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) + \rho g_y = 0 \tag{6.133}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{6.134}$$

---

[14]Since the flow is incompressible we will not distinguish between the strain rate tensor and the deviatoric strain rate tensor since they are identical.

Of course these PDEs must be supplemented by appropriate boundary conditions.

At this stage we must acknowledge (as in the FEM case) that it would be pointless to blindly discretise the equations above. Much work has been done on the topics for decades and only certain approaches have proven succesful. The book of Taras Gerya (1st and second edition [483, 484]) remains the reference in geodynamics and what follows is greatly influenced by this book.

6.12.2   A quick note about the MAC sheme

The marker and cell (MAC) scheme originates in Harlow and Welch [582] (1965) and is nicely summed up in Kanschat [709] (2008) and McKee et al. [900] (2008). Before going any further we need to highlight a few points from the original paper. First,

> One point should be emphasized. The marker particles introduced into this incompressible flow calculation are only for the purpose of indicating fluid configuration. They show which cells contain fluid and especially which cells lie along the free surface. [...] They do not participate in the calculation as do the Particle-in-Cell-method particles in the Los alamos computer program for compressible flows.

Second, looking at their equations

$$\frac{\partial u}{\partial t} + \frac{\partial u^2}{\partial x} + \frac{\partial uv}{\partial y} + \frac{\partial \varphi}{\partial x} = \nu \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + g_x,$$

$$\frac{\partial v}{\partial t} + \frac{\partial uv}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{\partial \varphi}{\partial y} = \nu \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + g_y.$$

we read that φ is the ratio of the pressure to the (constant) density and that $\nu = \eta/\rho$ is the kinematic viscosity coefficient. This constant-density approximation is quite limiting in the context of geodynamics where buoyancy-driven flow is important so we will not pursue this direction further. Likewise the constant-viscosity approximation is also not desirable.

Before moving on, it is quite interesting that when looking at their Fig. 2, we find that the spatial layout of the unknowns is of the staggered type:

FIG. 2. Field variable value placement about a computational cell. Velocities are defined at cell boundaries while pressures are defined at cell centers.
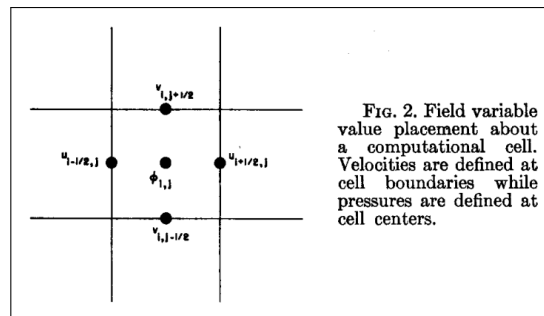
6.12.3   The fully staggered grid

Let us consider a 2D Cartesian domain of size $L_x \times L_y$. It can be shown that the following layout of nodes is necessary if the FDM is to be successful[15]: Note that in what follows we consider a mesh where all nodes are equidistant. This is not a requirement of the method but it allows for more compact notations[16].

---

[15]In other words any attempt where u, v, p would be on the corner nodes is doomed.

[16]We will revisit this topic in Section 6.12.15.
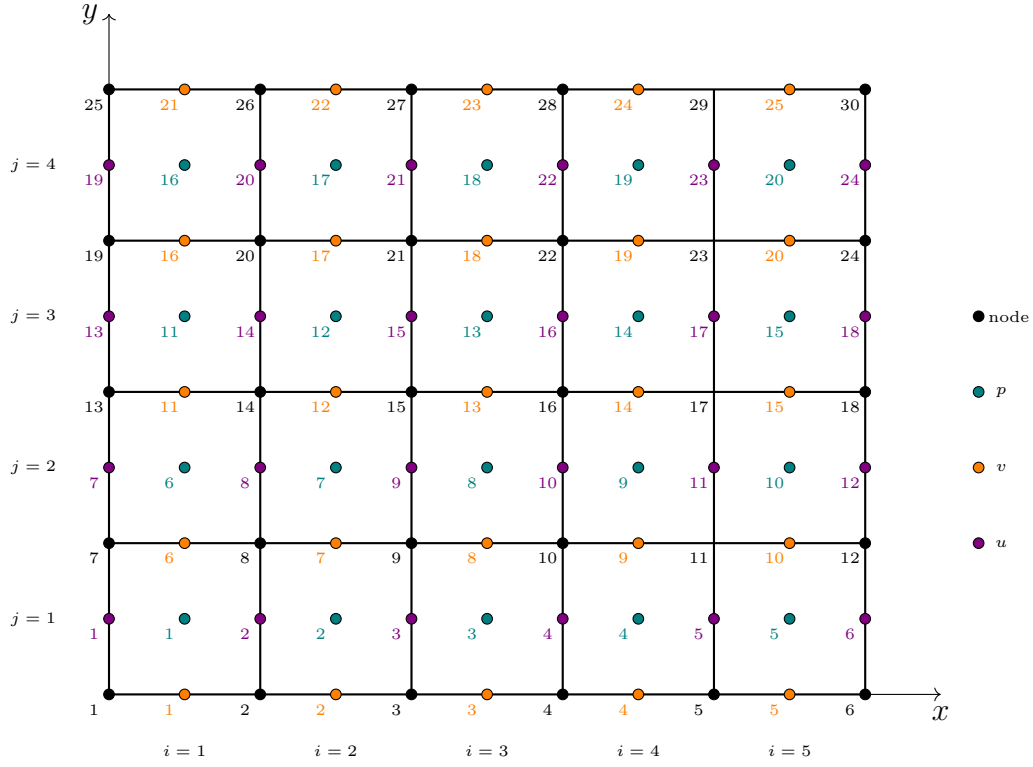
There are:

- $ncell = ncellx \cdot ncelly = ncell$ cells ($5 \cdot 4$ above)

- $N_p = ncell$ pressure values (20 above)

- $N_u = (ncellx + 1) \cdot ncelly$ u unknowns ($6 \cdot 4$ above)

- $N_p = ncell \cdot (ncelly + 1)$ v unknowns ($5 \cdot 5$ above)

- each cell is of size $h_x \cdot h_y$ with $h_x = L_x/ncellx$ and $h_y = L_y/ncelly$.

In total there are (not taking boundary conditions into account) N unknowns with:

$$N = N_u + N_v + N_p = (ncellx + 1) \cdot ncelly + ncell \cdot (ncelly + 1) + ncellx] \cdot ncelly$$

Note: the number of u unknowns is not necessarily equal to the number of v unknowns! This is nearly always the case with FEM.

This means that once we have discretised the Stokes equations we will need as many (i.e. N) equations in order to obtain a linear system with a solution.

We find a similar staggered grid figure in Gerya's book:

fully staggered 2D grid
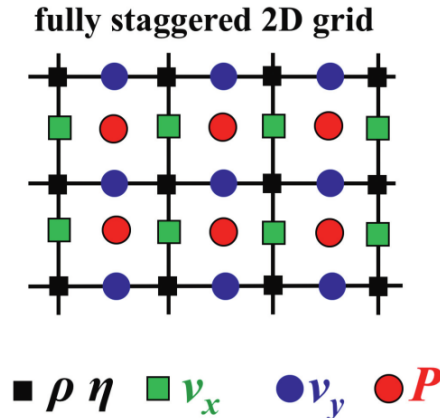
Taken from Gerya (2019).

What follows is mostly borrowed from Taras Gerya's book (second edition). However I do not like the notations of the book (or rather they do not match with the ones of this project) so I have changed them (in particular I use the 'n-e-w-s' system, i.e. north-east-west-south, and I prefer τ to σ' for the deviatoric stress tensor. Also I feel that it makes more sense to start with a regular mesh where all cells have the same size $h_x \times h_y$ (as opposed to the book). Finally, in the book the y-axis points downwards while 'my' y-axis points upwards.

In what follows we assume that we know the value of the density ρ and viscosity η on the (black) nodes of the mesh. We here only focus on the Stokes equations (mass and momentum conservation equations) and leave the energy equation aside.

The justification for the particular approach taken below to discretise the equations stems from the concept of conservative finite differences explained in Section 7.3 of Gerya's book.

6.12.4 Momentum equation: x-component

Let us zoom-in on a node which carries a u unknown. We label it $u_\otimes$. Its right ('east') neighbour is then u_e, its left ('west') neighbour is u_w, its top ('north') neighbour is u_n and its bottom ('south') neighbour is u_s. Its nearest pressure nodes are p_e and p_w, and the v nodes that surround it are v_{sw}, v_{se}, v_{ne} and v_{nw} as shown here:
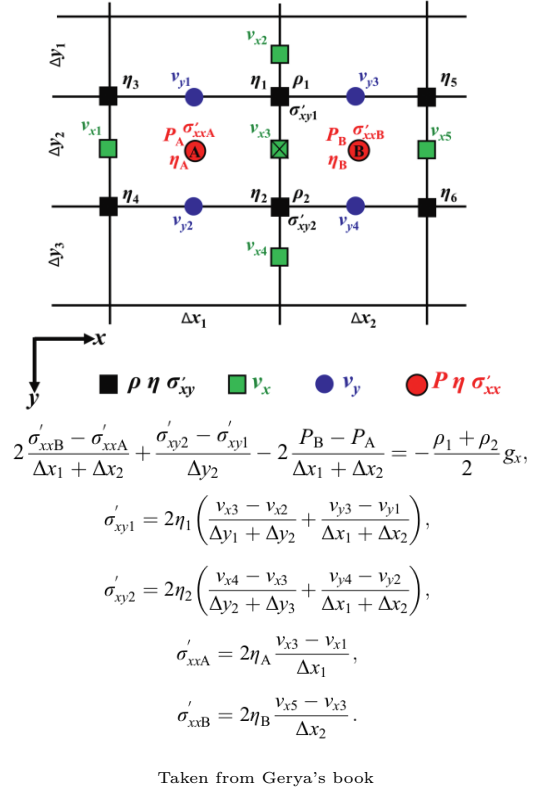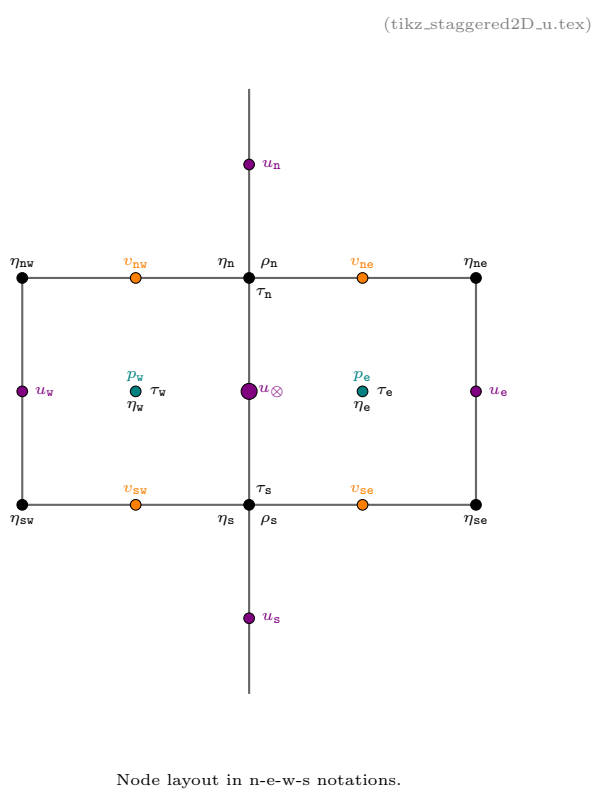
Node layout in n-e-w-s notations.

Taken from Gerya's book

We start from

$$\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} - \frac{\partial p}{\partial x} + \rho g_x = 0$$

which we discretise as follows on the purple node in the middle:

$$\frac{\tau_{xx,e} - \tau_{xx,w}}{h_x} + \frac{\tau_{xy,n} - \tau_{xy,s}}{h_y} - \frac{p_e - p_w}{h_x} = -\frac{\rho_n + \rho_s}{2} g_x$$

218

where

$$\tau_{xx,\mathsf{e}} = 2\eta_\mathsf{e}\dot{\varepsilon}_{xx,\mathsf{e}} = 2\eta_\mathsf{e}\frac{u_\mathsf{e} - u_\otimes}{h_x}$$

$$\tau_{xx,\mathsf{w}} = 2\eta_\mathsf{w}\dot{\varepsilon}_{xx,\mathsf{w}} = 2\eta_\mathsf{w}\frac{u_\otimes - u_\mathsf{w}}{h_x}$$

$$\tau_{xy,\mathsf{n}} = 2\eta_\mathsf{n}\dot{\varepsilon}_{xy,\mathsf{n}} = \eta_\mathsf{n}\left(\frac{u_\mathsf{n} - u_\otimes}{h_y} + \frac{v_\mathsf{ne} - v_\mathsf{nw}}{h_x}\right)$$

$$\tau_{xy,\mathsf{s}} = 2\eta_\mathsf{s}\dot{\varepsilon}_{xy,\mathsf{s}} = \eta_\mathsf{s}\left(\frac{u_\otimes - u_\mathsf{s}}{h_y} + \frac{v_\mathsf{se} - v_\mathsf{sw}}{h_x}\right)$$

We will discuss later in Section 6.12.7 how to arrive at η_e and η_w. Inserting the four equations above into the first one, we obtain:

$$\frac{1}{h_x}(\tau_{xx,\mathsf{e}} - \tau_{xx,\mathsf{w}}) + \frac{1}{h_y}(\tau_{xy,\mathsf{n}} - \tau_{xy,\mathsf{s}}) - \frac{1}{h_x}p_\mathsf{e} + \frac{1}{h_x}p_\mathsf{w} = -\frac{\rho_\mathsf{n} + \rho_\mathsf{s}}{2}g_x$$

$$\frac{1}{h_x}\left(2\eta_\mathsf{e}\frac{u_\mathsf{e} - u_\otimes}{h_x} - 2\eta_\mathsf{w}\frac{u_\otimes - u_\mathsf{w}}{h_x}\right) + \frac{1}{h_y}\left[\eta_\mathsf{n}\left(\frac{u_\mathsf{n} - u_\otimes}{h_y} + \frac{v_\mathsf{ne} - v_\mathsf{nw}}{h_x}\right) - \eta_\mathsf{s}\left(\frac{u_\otimes - u_\mathsf{s}}{h_y} + \frac{v_\mathsf{se} - v_\mathsf{sw}}{h_x}\right)\right] - \frac{1}{h_x}p_\mathsf{e} + \frac{1}{h_x}p_\mathsf{w} = -\frac{\rho_\mathsf{n} + \rho_\mathsf{s}}{2}g_x$$

$$\frac{2\eta_\mathsf{e}}{h_x}\frac{u_\mathsf{e} - u_\otimes}{h_x} - \frac{2\eta_\mathsf{w}}{h_x}\frac{u_\otimes - u_\mathsf{w}}{h_x} + \frac{\eta_\mathsf{n}}{h_y}\left(\frac{u_\mathsf{n} - u_\otimes}{h_y} + \frac{v_\mathsf{ne} - v_\mathsf{nw}}{h_x}\right) - \frac{\eta_\mathsf{s}}{h_y}\left(\frac{u_\otimes - u_\mathsf{s}}{h_y} + \frac{v_\mathsf{se} - v_\mathsf{sw}}{h_x}\right) - \frac{1}{h_x}p_\mathsf{e} + \frac{1}{h_x}p_\mathsf{w} = -\frac{\rho_\mathsf{n} + \rho_\mathsf{s}}{2}g_x$$

$$\begin{aligned}
&\left(\frac{\eta_\mathsf{n}}{h_y^2}\right)u_\mathsf{n} + \left(\frac{2\eta_\mathsf{e}}{h_x^2}\right)u_\mathsf{e} + \left(\frac{2\eta_\mathsf{w}}{h_x^2}\right)u_\mathsf{w} + \left(\frac{\eta_\mathsf{s}}{h_y^2}\right)u_\mathsf{s} + \left(-\frac{2\eta_\mathsf{e}}{h_x^2} - \frac{2\eta_\mathsf{w}}{h_x^2} - \frac{\eta_\mathsf{n}}{h_y^2} - \frac{\eta_\mathsf{s}}{h_y^2}\right)u_\otimes \\
&+ \left(\frac{\eta_\mathsf{n}}{h_x h_y}\right)v_\mathsf{ne} + \left(-\frac{\eta_\mathsf{n}}{h_x h_y}\right)v_\mathsf{nw} + \left(-\frac{\eta_\mathsf{s}}{h_x h_y}\right)v_\mathsf{se} + \left(\frac{\eta_\mathsf{s}}{h_x h_y}\right)v_\mathsf{sw} - \frac{1}{h_x}p_\mathsf{e} + \frac{1}{h_x}p_\mathsf{w} = -\frac{\rho_\mathsf{n} + \rho_\mathsf{s}}{2}g_x
\end{aligned} \tag{6.135}$$

219

6.12.5 Momentum equation: y-component
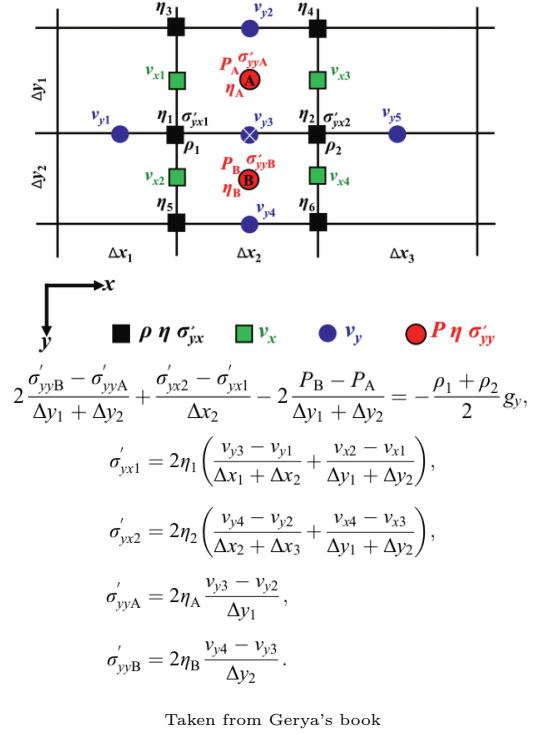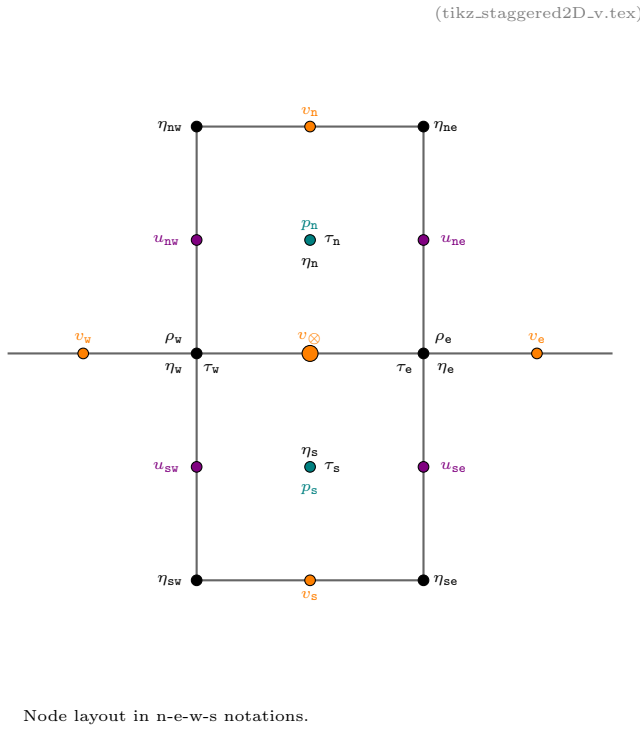
Node layout in n-e-w-s notations.

$$2\frac{\sigma'_{yyB} - \sigma'_{yyA}}{\Delta y_1 + \Delta y_2} + \frac{\sigma'_{yx2} - \sigma'_{yx1}}{\Delta x_2} - 2\frac{P_B - P_A}{\Delta y_1 + \Delta y_2} = -\frac{\rho_1 + \rho_2}{2}g_y,$$

$$\sigma'_{yx1} = 2\eta_1\left(\frac{v_{y3} - v_{y1}}{\Delta x_1 + \Delta x_2} + \frac{v_{x2} - v_{x1}}{\Delta y_1 + \Delta y_2}\right),$$

$$\sigma'_{yx2} = 2\eta_2\left(\frac{v_{y4} - v_{y2}}{\Delta x_2 + \Delta x_3} + \frac{v_{x4} - v_{x3}}{\Delta y_1 + \Delta y_2}\right),$$

$$\sigma'_{yyA} = 2\eta_A\frac{v_{y3} - v_{y2}}{\Delta y_1},$$

$$\sigma'_{yyB} = 2\eta_B\frac{v_{y4} - v_{y3}}{\Delta y_2}.$$

Taken from Gerya's book

We start from

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} - \frac{\partial p}{\partial y} + \rho g_y = 0$$

which we discretise as follows on the $v_\otimes$ node in the middle

$$\frac{\tau_{xy,\mathrm{e}} - \tau_{xy,\mathrm{w}}}{h_x} + \frac{\tau_{yy,\mathrm{n}} - \tau_{yy,\mathrm{s}}}{h_y} - \frac{p_\mathrm{n} - p_\mathrm{s}}{h_y} = -\frac{\rho_\mathrm{e} + \rho_\mathrm{w}}{2}g_y$$

where

$$\tau_{xy,\mathrm{e}} = 2\eta_\mathrm{e}\dot{\varepsilon}_{xy,\mathrm{e}} = \eta_\mathrm{e}\left(\frac{u_\mathrm{ne} - u_\mathrm{se}}{h_y} + \frac{v_\mathrm{e} - v_\otimes}{h_x}\right)$$

$$\tau_{xy,\mathrm{w}} = 2\eta_\mathrm{w}\dot{\varepsilon}_{xy,\mathrm{w}} = \eta_\mathrm{w}\left(\frac{u_\mathrm{nw} - u_\mathrm{sw}}{h_y} + \frac{v_\otimes - v_\mathrm{w}}{h_x}\right)$$

$$\tau_{yy,\mathrm{n}} = 2\eta_\mathrm{n}\dot{\varepsilon}_{yy,\mathrm{n}} = 2\eta_\mathrm{n}\frac{v_\mathrm{n} - v_\otimes}{h_y}$$

$$\tau_{yy,\mathrm{s}} = 2\eta_\mathrm{s}\dot{\varepsilon}_{yy,\mathrm{s}} = 2\eta_\mathrm{s}\frac{v_\otimes - v_\mathrm{s}}{h_y}$$

Again, we will discuss later in Section 6.12.7 how to arrive at η_n and η_s. Inserting the four equations above into the first one, we obtain:

$$\frac{1}{h_x}\left(\tau_{xy,\mathrm{e}} - \tau_{xy,\mathrm{w}}\right) + \frac{1}{h_y}\left(\tau_{yy,\mathrm{n}} - \tau_{yy,\mathrm{s}}\right) - \frac{1}{h_y}p_\mathrm{n} + \frac{1}{h_y}p_\mathrm{s} = -\frac{\rho_\mathrm{e} + \rho_\mathrm{w}}{2}g_y$$

$$\frac{1}{h_x}\left[\eta_\mathrm{e}\left(\frac{u_\mathrm{ne} - u_\mathrm{se}}{h_y} + \frac{v_\mathrm{e} - v_\otimes}{h_x}\right) - \eta_\mathrm{w}\left(\frac{u_\mathrm{nw} - u_\mathrm{sw}}{h_y} + \frac{v_\otimes - v_\mathrm{w}}{h_x}\right)\right] + \frac{1}{h_y}\left(2\eta_\mathrm{n}\frac{v_\mathrm{n} - v_\otimes}{h_y} - 2\eta_\mathrm{s}\frac{v_\otimes - v_\mathrm{s}}{h_y}\right) - \frac{1}{h_y}p_\mathrm{n} + \frac{1}{h_y}p_\mathrm{s} = -\frac{\rho_\mathrm{e} + \rho_\mathrm{w}}{2}g_y$$

$$\frac{\eta_\mathrm{e}}{h_x}\left(\frac{u_\mathrm{ne} - u_\mathrm{se}}{h_y} + \frac{v_\mathrm{e} - v_\otimes}{h_x}\right) - \frac{\eta_\mathrm{w}}{h_x}\left(\frac{u_\mathrm{nw} - u_\mathrm{sw}}{h_y} + \frac{v_\otimes - v_\mathrm{w}}{h_x}\right) + \frac{2\eta_\mathrm{n}}{h_y}\frac{v_\mathrm{n} - v_\otimes}{h_y} - \frac{2\eta_\mathrm{s}}{h_y}\frac{v_\otimes - v_\mathrm{s}}{h_y} - \frac{1}{h_y}p_\mathrm{n} + \frac{1}{h_y}p_\mathrm{s} = -\frac{\rho_\mathrm{e} + \rho_\mathrm{w}}{2}g_y$$

$$\left(\frac{2\eta_\mathtt{n}}{h_y^2}\right) v_\mathtt{n} + \left(\frac{\eta_\mathtt{e}}{h_x^2}\right) v_\mathtt{e} + \left(\frac{\eta_\mathtt{w}}{h_x^2}\right) v_\mathtt{w} + \left(\frac{2\eta_\mathtt{s}}{h_y^2}\right) v_\mathtt{s} + \left(-\frac{\eta_\mathtt{e}}{h_x^2} - \frac{\eta_\mathtt{w}}{h_x^2} - \frac{2\eta_\mathtt{n}}{h_y^2} - \frac{2\eta_\mathtt{s}}{h_y^2}\right) v_\otimes$$
$$+ \left(\frac{\eta_\mathtt{e}}{h_x h_y}\right) u_\mathtt{ne} + \left(-\frac{\eta_\mathtt{e}}{h_x h_y}\right) u_\mathtt{se} + \left(-\frac{\eta_\mathtt{w}}{h_x h_y}\right) u_\mathtt{nw} + \left(\frac{\eta_\mathtt{w}}{h_x h_y}\right) u_\mathtt{sw} - \frac{1}{h_y} p_\mathtt{n} + \frac{1}{h_y} p_\mathtt{s} \;\; = \;\; -\frac{\rho_\mathtt{e} + \rho_\mathtt{w}}{2} g_y$$
$$(6.136)$$

6.12.6 Continuity equation

In two dimensions the continuity equation for an incompressible flow is

$$\vec{\nabla} \cdot \vec{\mathbf{v}} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

We can isolate one cell:
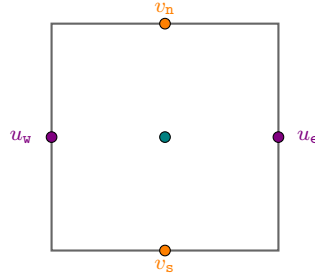
and discretise the equation above in its middle.

$$\frac{u_\mathtt{e} - u_\mathtt{w}}{h_x} + \frac{v_\mathtt{n} - v_\mathtt{s}}{h_y} = 0 \qquad\qquad (6.137)$$

6.12.7 Viscosity at the center of cells

Following Gerya, the viscosity in the middle of the cells can be obtained via averaging (for example arithmetic) of the viscosity values at the corners.

In the case of the x-component of the momentum equation:

$$\eta_\mathtt{w} = \frac{1}{4}(\eta_\mathtt{sw} + \eta_\mathtt{s} + \eta_\mathtt{n} + \eta_\mathtt{nw})$$

$$\eta_\mathtt{e} = \frac{1}{4}(\eta_\mathtt{se} + \eta_\mathtt{s} + \eta_\mathtt{n} + \eta_\mathtt{ne})$$

In the case of the y-component of the momentum equation:

$$\eta_\mathtt{n} = \frac{1}{4}(\eta_\mathtt{w} + \eta_\mathtt{nw} + \eta_\mathtt{ne} + \eta_\mathtt{e})$$

$$\eta_\mathtt{s} = \frac{1}{4}(\eta_\mathtt{sw} + \eta_\mathtt{w} + \eta_\mathtt{e} + \eta_\mathtt{se})$$

If harmonic averaging is preferred, then for instance we would have

$$\eta_\mathtt{w} = \frac{4}{1/\eta_\mathtt{sw} + 1/\eta_\mathtt{s} + 1/\eta_\mathtt{n} + 1/\eta_\mathtt{nw}}$$

If geometric averaging is preferred, then for instance we would have

$$\log_{10} \eta_\mathtt{s} = \frac{1}{4}(\log_{10} \eta_\mathtt{sw} + \log_{10} \eta_\mathtt{w} + \log_{10} \eta_\mathtt{e} + \log_{10} \eta_\mathtt{se})$$

6.12.8   Boundary conditions

We start with so-called Dirichlet boundary conditions, i.e. we wish to prescribe the value of all/some of velocity components on the boundaries. For simplicity we will then focus on free slip and no slip boundary conditions. Free slip b.c. are characterised by a) the normal velocity component on the boundary is zero b) the other two components do not change across the boundary (i.e. zero shear strain rate and shear stresses along the boundary). No slip b.c. require that all velocity components are zero on the boundary.

Gerya presents in Section 7.4 all the types of boundary conditions but is remarkably elusive as to how they should be implemented. Consulting Becker and Kaus, the same conclusion can unfortunately be drawn.

Looking at Harlow and Welch [582] (1965) we see that the authors discuss free-slip and no-slip boundary conditions:

Fig. 3. Field variables near a wall.
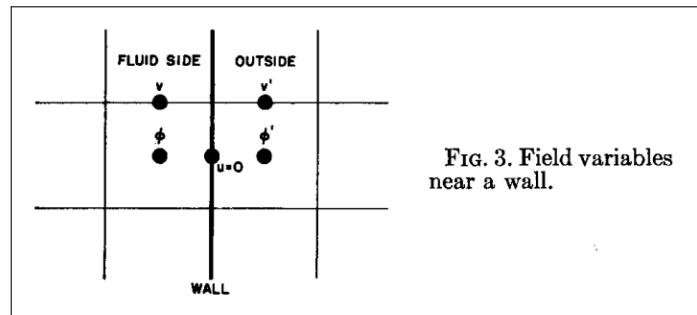
As mentioned in the paper "a vertical wall therefore passes through the horizontal-velocity mesh points, and the velocities at those points vanish at all times for either type of [boundary condition]. A vertical wall does not pass through vertical-velocity mesh points, but the calculation makes use of the values of v at [virtual] mesh points lying just outside of the wall. For a no-slip wall the boundary condition is $v' = -v$, while for a free-slip wall it becomes $v' = v$ [on the figure above]."

Let us focus on the bottom of the following mesh:
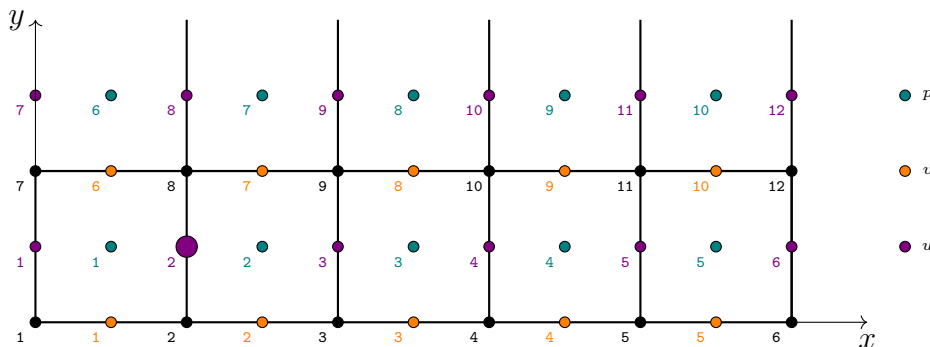
- No slip (at the bottom). We then want $u = v = 0$. We can easily zero $v_{1,2,3,4,5}$ since the corresponding nodes are on the boundary but this is not enough. Let us consider node u_2 (bigger on the figure above). We could zero u_2 directly but this would not be correct since the node is not on the boundary. If we wish to impose $u = 0$ exactly on the boundary then its south neighbour ghost value should be set to $-u_2$ and used in the established stencils.

Free slip   No slip
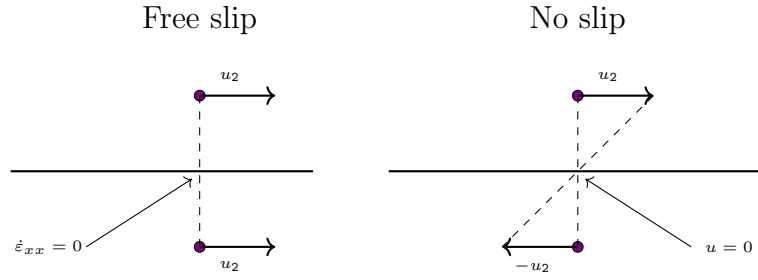
- <u>Free slip</u> (at the bottom). Aside from the zero normal velocity as above, we also need the tangential velocity gradient to be zero on the boundary, which translates into assigning its south neighbour ghost the value $+u_2$

We see that in both cases we will need to create ephemeral boundary nodes outside the domain and assign them specific values based on the boundary condition type. This consideration led Becker & Kaus to produce this figure in their syllabus:
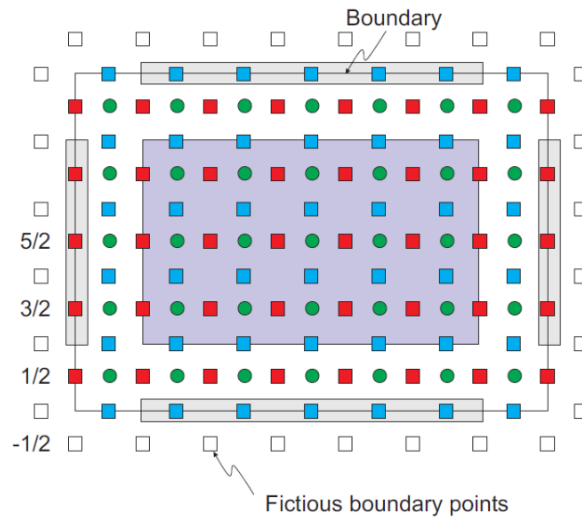
Taken from Becker and Kaus [71]: Staggered grid definition with the boundary points. Within the purple domain, the finite difference scheme for center points can be applied. At the boundaries, we have to apply a special finite difference scheme which employs fictitious boundary nodes.

Let us consider a 4×3 mesh. We have 12 cells, 15 u-nodes, 16 v-nodes and 12 p-nodes. This mesh then counts $N = 15 + 16 + 12 = 43$ unknowns[17].

(tikz_staggered2D_4x3.tex)

<hr>
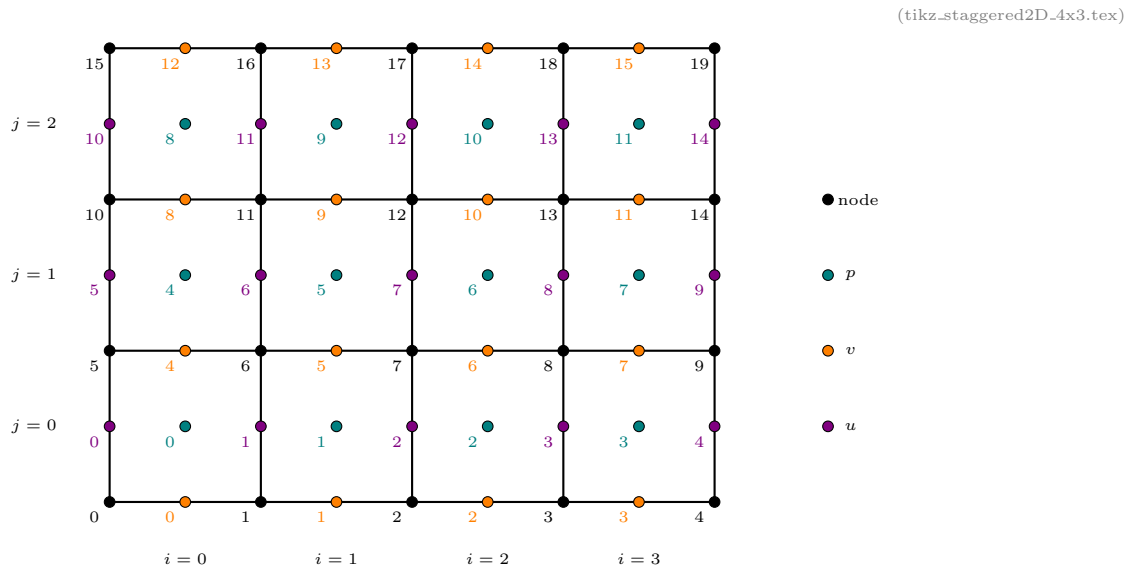
[17]Note that the numbering starts at zero, python-style.

Let us assume that we prescribe free slip or no slip on all sides. Then we must set $v_0 = v_1 = v_2 = v_3 = v_{12} = v_{13} = v_{14} = v_{15} = 0$ and $u_0 = u_4 = u_5 = u_9 = u_{10} = u_{14} = 0$. This means that we need not discretise the momentum equation on these nodes.

Based on the considerations above pertaining to ghost/virtual nodes we turn once more to our 4×3 cell mesh where ghost/virtual nodes have been added:

- $u_{0,5,10}$: situated on left boundary. Free slip and no slip require $u = 0$. No need for a ghost node.

- $u_{4,9,14}$: situated on right boundary. Free slip and no slip require $u = 0$. No need for a ghost node.

- $u_{1,2,3}$: north, east, west neighbours are present, but south is missing so south ghost nodes are needed.

Let us write the stencil for these nodes:

$$\left(\frac{\eta_\mathtt{n}}{h_y^2}\right) u_\mathtt{n} + \left(\frac{2\eta_\mathtt{e}}{h_x^2}\right) u_\mathtt{e} + \left(\frac{2\eta_\mathtt{w}}{h_x^2}\right) u_\mathtt{w} + \boxed{\left(\frac{\eta_\mathtt{s}}{h_y^2}\right) u_\mathtt{s}} + \left(-\frac{2\eta_\mathtt{e}}{h_x^2} - \frac{2\eta_\mathtt{w}}{h_x^2} - \frac{\eta_\mathtt{n}}{h_y^2} - \frac{\eta_\mathtt{s}}{h_y^2}\right) u_\otimes$$

$$+ \left(\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_\mathtt{ne} + \left(-\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_\mathtt{nw} + \left(-\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_\mathtt{se} + \left(\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_\mathtt{sw} - \frac{1}{h_x} p_\mathtt{e} + \frac{1}{h_x} p_\mathtt{w} = -\frac{\rho_\mathtt{n} + \rho_\mathtt{s}}{2} g_x$$

As we have seen earlier there is obviously a problem as $u_\mathtt{s}$ (the boxed term) is not defined. If we wish to prescribe free slip, we have seen that we must set $u_\mathtt{s} = u_\otimes$. If we wish to prescribe no slip, we have seen that we must set $u_\mathtt{s} = -u_\otimes$. The stencil then becomes:

$$\left(\frac{\eta_\mathtt{n}}{h_y^2}\right) u_\mathtt{n} + \left(\frac{2\eta_\mathtt{e}}{h_x^2}\right) u_\mathtt{e} + \left(\frac{2\eta_\mathtt{w}}{h_x^2}\right) u_\mathtt{w} + \left(-\frac{2\eta_\mathtt{e}}{h_x^2} - \frac{2\eta_\mathtt{w}}{h_x^2} - \frac{\eta_\mathtt{n}}{h_y^2} - (1 - \delta_{bc})\frac{\eta_\mathtt{s}}{h_y^2}\right) u_\otimes$$

$$+ \left(\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_\mathtt{ne} + \left(-\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_\mathtt{nw} + \left(-\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_\mathtt{se} + \left(\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_\mathtt{sw} - \frac{1}{h_x} p_\mathtt{e} + \frac{1}{h_x} p_\mathtt{w} = -\frac{\rho_\mathtt{n} + \rho_\mathtt{s}}{2} g_x$$

where $\delta_{bc} = 1$ (free slip) or -1 (no slip).

- $u_{11,12,13}$: south, east, west neighbours are present, but north is missing so north ghost nodes are needed.

  Let us write the stencil for these nodes:

  $$\boxed{\left(\frac{\eta_{\mathrm{n}}}{h_y^2}\right) u_{\mathrm{n}}} + \left(\frac{2\eta_{\mathrm{e}}}{h_x^2}\right) u_{\mathrm{e}} + \left(\frac{2\eta_{\mathrm{w}}}{h_x^2}\right) u_{\mathrm{w}} + \left(\frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{\mathrm{s}} + \left(-\frac{2\eta_{\mathrm{e}}}{h_x^2} - \frac{2\eta_{\mathrm{w}}}{h_x^2} - \frac{\eta_{\mathrm{n}}}{h_y^2} - \frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{\otimes}$$

  $$+ \left(\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{\mathrm{ne}} + \left(-\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{\mathrm{nw}} + \left(-\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_{\mathrm{se}} + \left(\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_{\mathrm{sw}} - \frac{1}{h_x} p_{\mathrm{e}} + \frac{1}{h_x} p_{\mathrm{w}} \quad = \quad -\frac{\rho_{\mathrm{n}} + \rho_{\mathrm{s}}}{2} g_x$$

  Here again there is obviously a problem as u_{n} is not defined (boxed term). Using the same δ_{bc} parameter the stencil then becomes:

  $$\left(\frac{2\eta_{\mathrm{e}}}{h_x^2}\right) u_{\mathrm{e}} + \left(\frac{2\eta_{\mathrm{w}}}{h_x^2}\right) u_{\mathrm{w}} + \left(\frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{\mathrm{s}} + \left(-\frac{2\eta_{\mathrm{e}}}{h_x^2} - \frac{2\eta_{\mathrm{w}}}{h_x^2} - (1 - \delta_{bc})\frac{\eta_{\mathrm{n}}}{h_y^2} - \frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{\otimes}$$

  $$+ \left(\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{\mathrm{ne}} + \left(-\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{\mathrm{nw}} + \left(-\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_{\mathrm{se}} + \left(\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_{\mathrm{sw}} - \frac{1}{h_x} p_{\mathrm{e}} + \frac{1}{h_x} p_{\mathrm{w}} \quad = \quad -\frac{\rho_{\mathrm{n}} + \rho_{\mathrm{s}}}{2} g_x$$

- $u_{6,7,8}$: standard stencil applies as these nodes have four 'real' neighbours. Let us consider node u_7. The x-component of the momentum equation there is written:

  $$\left(\frac{\eta_{\mathrm{n}}}{h_y^2}\right) u_{12} + \left(\frac{2\eta_{\mathrm{e}}}{h_x^2}\right) u_8 + \left(\frac{2\eta_{\mathrm{w}}}{h_x^2}\right) u_5 + \left(\frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_2 + \left(-\frac{2\eta_{\mathrm{e}}}{h_x^2} - \frac{2\eta_{\mathrm{w}}}{h_x^2} - \frac{\eta_{\mathrm{n}}}{h_y^2} - \frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_7$$

  $$+ \left(\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{10} + \left(-\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_9 + \left(-\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_6 + \left(\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_5 - \frac{1}{h_x} p_6 + \frac{1}{h_x} p_5 = -\frac{\rho_{12} + \rho_7}{2} g_x$$

- $v_{0,1,2,3}$: situated on bottom boundary. Free slip and no slip require $v = 0$. No need for a ghost node.

- $v_{12,13,14,15}$: situated on top boundary. Free slip and no slip require $v = 0$. No need for a ghost node.

- $v_{4,8}$: north, east, south neighbours are present, but west is missing so west ghost nodes are needed.

  $$\left(\frac{2\eta_{\mathrm{n}}}{h_y^2}\right) v_{\mathrm{n}} + \left(\frac{\eta_{\mathrm{e}}}{h_x^2}\right) v_{\mathrm{e}} + \boxed{\left(\frac{\eta_{\mathrm{w}}}{h_x^2}\right) v_{\mathrm{w}}} + \left(\frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\mathrm{s}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x^2} - \frac{\eta_{\mathrm{w}}}{h_x^2} - \frac{2\eta_{\mathrm{n}}}{h_y^2} - \frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\otimes}$$

  $$+ \left(\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{ne}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{se}} + \left(-\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{nw}} + \left(\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{sw}} - \frac{1}{h_y} p_{\mathrm{n}} + \frac{1}{h_y} p_{\mathrm{s}} \quad = \quad -\frac{\rho_{\mathrm{e}} + \rho_{\mathrm{w}}}{2} g_y$$

  leads to

  $$\left(\frac{2\eta_{\mathrm{n}}}{h_y^2}\right) v_{\mathrm{n}} + \left(\frac{\eta_{\mathrm{e}}}{h_x^2}\right) v_{\mathrm{e}} + \left(\frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\mathrm{s}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x^2} - (1 - \delta_{bc})\frac{\eta_{\mathrm{w}}}{h_x^2} - \frac{2\eta_{\mathrm{n}}}{h_y^2} - \frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\otimes}$$

  $$+ \left(\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{ne}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{se}} + \left(-\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{nw}} + \left(\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{sw}} - \frac{1}{h_y} p_{\mathrm{n}} + \frac{1}{h_y} p_{\mathrm{s}} \quad = \quad -\frac{\rho_{\mathrm{e}} + \rho_{\mathrm{w}}}{2} g_y$$

- $v_{7,11}$: south, west, north neighbours are present, but east is missing so east ghost nodes are needed.

  $$\left(\frac{2\eta_{\mathrm{n}}}{h_y^2}\right) v_{\mathrm{n}} + \boxed{\left(\frac{\eta_{\mathrm{e}}}{h_x^2}\right) v_{\mathrm{e}}} + \left(\frac{\eta_{\mathrm{w}}}{h_x^2}\right) v_{\mathrm{w}} + \left(\frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\mathrm{s}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x^2} - \frac{\eta_{\mathrm{w}}}{h_x^2} - \frac{2\eta_{\mathrm{n}}}{h_y^2} - \frac{2\eta_{\mathrm{s}}}{h_y^2}\right) v_{\otimes}$$

  $$+ \left(\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{ne}} + \left(-\frac{\eta_{\mathrm{e}}}{h_x h_y}\right) u_{\mathrm{se}} + \left(-\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{nw}} + \left(\frac{\eta_{\mathrm{w}}}{h_x h_y}\right) u_{\mathrm{sw}} - \frac{1}{h_y} p_{\mathrm{n}} + \frac{1}{h_y} p_{\mathrm{s}} \quad = \quad -\frac{\rho_{\mathrm{e}} + \rho_{\mathrm{w}}}{2} g_y$$

leads to

$$\left(\frac{2\eta_{\mathtt{n}}}{h_y^2}\right) v_{\mathtt{n}} + \left(\frac{\eta_{\mathtt{w}}}{h_x^2}\right) v_{\mathtt{w}} + \left(\frac{2\eta_{\mathtt{s}}}{h_y^2}\right) v_{\mathtt{s}} + \left(-(1-\delta_{bc})\frac{\eta_{\mathtt{e}}}{h_x^2} - \frac{\eta_{\mathtt{w}}}{h_x^2} - \frac{2\eta_{\mathtt{n}}}{h_y^2} - \frac{2\eta_{\mathtt{s}}}{h_y^2}\right) v_{\otimes}$$
$$+ \left(\frac{\eta_{\mathtt{e}}}{h_x h_y}\right) u_{\mathtt{ne}} + \left(-\frac{\eta_{\mathtt{e}}}{h_x h_y}\right) u_{\mathtt{se}} + \left(-\frac{\eta_{\mathtt{w}}}{h_x h_y}\right) u_{\mathtt{nw}} + \left(\frac{\eta_{\mathtt{w}}}{h_x h_y}\right) u_{\mathtt{sw}} - \frac{1}{h_y}p_{\mathtt{n}} + \frac{1}{h_y}p_{\mathtt{s}} \quad = \quad -\frac{\rho_{\mathtt{e}} + \rho_{\mathtt{w}}}{2}g_y$$

- $v_{5,6,9,10}$: standard stencil applies as these nodes have four 'real' neighbours.

- pressure nodes are situated inside each cell and are unaffected by the type of boundary conditions since the stencil of the continuity equation is always complete (no need for ghost nodes).

Note that if free slip or no slip boundary conditions are prescribed on all four boundaries the pressure solution is only known up to a constant. Often this pressure nullspace is removed from the solution since most solvers will happily solve the linear system even though it is not strictly definite. Gerya advocates to set a zero boundary condition on a node (and then re-normalise the pressure).

In practice ghost nodes are never created or used. They are useful in order to establish the corrected stencils but are entirely absent from the coding.

Note that Duretz, May, Gerya, and Tackley [375] (2011) calls the method above the "fictitious node method" and states:

> This method do not require any additional pressure constraints, thus no boundary conditions are required while solving for pressure. With this discretization of the boundary condition, each partial derivative is approximated by means of central finite differences and we therefore expect this method to be second-order accurate in space.

The authors also present the "extrapolated BC method" but state that it must be supplemented with pressure conditions and that

> Given the fact that the boundary condition is defined via extrapolation, we expect this implementation to be first order in space.

This is later proven, see Fig. 7 of the paper.

6.12.9 Generating the linear system

Let us now create the vector $\vec{X}$ that is $N = N_u + N_v + N_p = 43$ long (still considering the 4×3 cell mesh presented above):

$$\vec{X}^T = (\vec{\mathcal{U}}, \vec{\mathcal{V}}, \vec{\mathcal{P}}) = (u_0, u_1, \ldots, u_{14}, v_0, v_1, \ldots, v_{15}, p_0, p_1, \ldots, p_{11})$$

We have 14 boundary conditions and 29 stencil equations that establish relationships between unknowns (see previous section) which we can write altogether as a linear system.

The boundary condition $u_0 = 0$ can be written

$$(1, 0, 0, 0, \ldots, 0) \cdot \vec{X} = (0, \ldots)^T$$

so that the lhs horizontal vector becomes the first line of the matrix and we must write zero in the first element of the rhs vector.

Likewise the boundary condition $u_4 = 0$ can be written

$$(0, 0, 0, 0, 1, \ldots, 0) \cdot \vec{X} = (., ., ., ., 0, \ldots)^T$$

and the lhs horizontal vector becomes the 5th line of the matrix and we must write zero in the 5th element of the rhs vector.

We repeat this procedure for all u boundary conditions, which allows us to fill the 1st, 5th, 6th, 10th, 11th and 14th line of the matrix.

Likewise the boundary conditions on v will yield similar lines in the matrix, only the corresponding lines are below the 15 first ones corresponding to u.

Considering node u_7, the stencil equation (6.135) then becomes:

$$\left(\frac{\eta_\mathtt{n}}{h_y^2}\right) u_{12} + \left(\frac{2\eta_\mathtt{e}}{h_x^2}\right) u_8 + \left(\frac{2\eta_\mathtt{w}}{h_x^2}\right) u_6 + \left(\frac{\eta_\mathtt{s}}{h_y^2}\right) u_2 + \left(-\frac{2\eta_\mathtt{e}}{h_x^2} - \frac{2\eta_\mathtt{w}}{h_x^2} - \frac{\eta_\mathtt{n}}{h_y^2} - \frac{\eta_\mathtt{s}}{h_y^2}\right) u_7$$

$$+ \left(\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_{10} + \left(-\frac{\eta_\mathtt{n}}{h_x h_y}\right) v_9 + \left(-\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_6 + \left(\frac{\eta_\mathtt{s}}{h_x h_y}\right) v_5 - \frac{1}{h_x} p_6 + \frac{1}{h_x} p_5 \quad = \quad -\frac{\rho_{12} + \rho_7}{2} g_x$$

which will correspond to the 7th line in the matrix. The coefficients inside the parentheses will be input in the column corresponding to the unknowns they front.

Because of how the $\vec{X}$ vector is built (first u, then v then p unknowns) the resulting linear system takes the following form:

$$\begin{pmatrix} \mathbb{K}_{xx} & \mathbb{K}_{xy} & \mathbb{G}_x \\ \mathbb{K}_{yx} & \mathbb{K}_{yy} & \mathbb{G}_y \\ \mathbb{Q}_x & \mathbb{Q}_y & 0 \end{pmatrix} \cdot \vec{X} = \vec{b} \tag{6.138}$$

6.12.10 A remark about scaling of terms

The coefficients before the velocity components are all of the form η/h^2 while those in front of the pressure terms are of the form $1/h$.

In geodynamics typical velocities are of the order of a cm/year, or about 10^{-10} m/s while typical pressures are on the order of 1 GPa. For a viscosity of about 10^{21}Pa, and a mesh size of about 1km, we have:

$$\frac{\eta}{h^2} v \sim \frac{10^{21}}{1000^2} 10^{-10} \sim 10^6$$

while

$$\frac{p}{h} \sim \frac{10^9}{1000} = 10^6$$

All is fine until we realise that the terms η/h^2 and $1/h$ will form the entries of the matrix, with $\eta/h^2 \sim 10^{15}$ and $1/h \sim 10^{-3}$. These numbers are more than 10 orders of magnitude apart and will lead to numerical inaccuracies inside solvers (direct or iterative). It is therefore common[18] to define $p^\star = \frac{L_{ref}}{\eta_{ref}} p$ so that for example the y-momentum equation stencil becomes

$$\left(\frac{2\eta_\mathtt{n}}{h_y^2}\right) v_\mathtt{n} + \left(\frac{\eta_\mathtt{e}}{h_x^2}\right) v_\mathtt{e} + \left(\frac{\eta_\mathtt{w}}{h_x^2}\right) v_\mathtt{w} + \left(\frac{2\eta_\mathtt{s}}{h_y^2}\right) v_\mathtt{s} + \left(-\frac{\eta_\mathtt{e}}{h_x^2} - \frac{\eta_\mathtt{w}}{h_x^2} - \frac{2\eta_\mathtt{n}}{h_y^2} - \frac{2\eta_\mathtt{s}}{h_y^2}\right) v_\otimes$$

$$+ \left(\frac{\eta_\mathtt{e}}{h_x h_y}\right) u_\mathtt{ne} + \left(-\frac{\eta_\mathtt{e}}{h_x h_y}\right) u_\mathtt{se} + \left(-\frac{\eta_\mathtt{w}}{h_x h_y}\right) u_\mathtt{nw} + \left(\frac{\eta_\mathtt{w}}{h_x h_y}\right) u_\mathtt{sw} - \frac{1}{h_y} \frac{\eta_{ref}}{L_{ref}} p_\mathtt{n}^\star + \frac{1}{h_y} \frac{\eta_{ref}}{L_{ref}} p_\mathtt{s}^\star \quad = \quad -\frac{\rho_\mathtt{e} + \rho_\mathtt{w}}{2} g_y$$

By appropriately choosing the reference viscosity η_{ref} and the reference length L_{ref} we thereby ensure that the resulting matrix contains coefficients which are all only a few orders of magnitude apart (reflecting the viscosity range in the domain).

In the end the linear system will take the form

---

[18]See entry on the same topic in the FEM section.

$$\begin{pmatrix} \mathbb{K}_{xx} & \mathbb{K}_{xy} & \frac{\eta_{ref}}{L_{ref}}\mathbb{G}_x \\ \mathbb{K}_{yx} & \mathbb{K}_{yy} & \frac{\eta_{ref}}{L_{ref}}\mathbb{G}_y \\ \frac{\eta_{ref}}{L_{ref}}\mathbb{Q}_x & \frac{\eta_{ref}}{L_{ref}}\mathbb{Q}_y & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{U}} \\ \vec{\mathcal{V}} \\ \vec{\mathcal{P}}^{\star} \end{pmatrix} = \begin{pmatrix} \vec{b}_x \\ \vec{b}_y \\ \vec{0} \end{pmatrix} \quad\quad (6.139)$$

6.12.11   Making matrix symmetric again

Looking at Eq. (6.138) we see that the matrix is not structurally symmetric. In other words $\mathbb{K}_{xy} \neq \mathbb{K}_{yx}$, $\mathbb{Q}_x \neq \mathbb{G}_x^T$, and $\mathbb{Q}_y \neq \mathbb{G}_y^T$, which is obvious only looking at the non zero terms on the figures above.

This stems from how boundary conditions are applied. So far the matrix is being built line by line. Each line corresponding to a node where the value is prescribed on the boundary receives a 1 on the diagonal and the required value (often simply zero) in the rhs.

This is for example happening on the line corresponding to node u_{10}. But if we then turn to node u_{11}, the stencil equation on this node reads

$$\left(\frac{\eta_{\mathrm{n}}}{h_y^2}\right) u_{\mathrm{n}} + \left(\frac{2\eta_{\mathrm{e}}}{h_x^2}\right) u_{12} + \left(\frac{2\eta_{\mathrm{w}}}{h_x^2}\right) u_{10} + \left(\frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_6 + \left(-\frac{2\eta_{\mathrm{e}}}{h_x^2} - \frac{2\eta_{\mathrm{w}}}{h_x^2} - \frac{\eta_{\mathrm{n}}}{h_y^2} - \frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{11}$$

$$+ \left(\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{13} + \left(-\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{12} + \left(-\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_9 + \left(\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_8 - \frac{1}{h_x}p_9 + \frac{1}{h_x}p_8 \;\;=\;\; -\frac{\rho_{16} + \rho_{11}}{2}g_x$$

We have of course already seen how to deal with the missing north node which has us write

$$\left(\frac{2\eta_{\mathrm{e}}}{h_x^2}\right) u_{12} + \boxed{\left(\frac{2\eta_{\mathrm{w}}}{h_x^2}\right) u_{10}} + \left(\frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_6 + \left(-\frac{2\eta_{\mathrm{e}}}{h_x^2} - \frac{2\eta_{\mathrm{w}}}{h_x^2} - (1-\delta_{bc})\frac{\eta_{\mathrm{n}}}{h_y^2} - \frac{\eta_{\mathrm{s}}}{h_y^2}\right) u_{11}$$

$$+ \boxed{\left(\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{13}} + \boxed{\left(-\frac{\eta_{\mathrm{n}}}{h_x h_y}\right) v_{12}} + \left(-\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_9 + \left(\frac{\eta_{\mathrm{s}}}{h_x h_y}\right) v_8 - \frac{1}{h_x}p_9 + \frac{1}{h_x}p_8 \;\;=\;\; -\frac{\rho_{16} + \rho_{11}}{2}g_x$$

However, the boxed terms above are *also* on the boundary, which means that the values of u_{10}, v_{12} and v_{13} are known (typically zero). As such these two terms should make their way to the right hand side, thereby removing entries in the matrix and re-establishing the symmetry of the matrix. Note that it is not wrong to leave these entries, they just do not serve any real purpose.

6.12.12   Recovering the strain rate tensor from the velocity solution

Having obtained the velocity field inside the domain we may now wish to compute the components of the strain rate tensor, either for plotting purposes or because (for example) the effective viscosity depends on it.

If the Particle-In-Cell technique is used the strain rate needs to be computed on each particle. I need to check how FDM people do this.

If it is for plotting purposes, a quick look at the the discretisation of the continuity equation tells us that we know how to compute $\dot\varepsilon_{xx} = \partial u/\partial x$ and $\dot\varepsilon_{yy} = \partial v/\partial y$ in the middle of each cell. But what about $\dot\varepsilon_{xy}$?

Let us put the two node layouts we have used previously in the context of the x- and y-momentum equation:
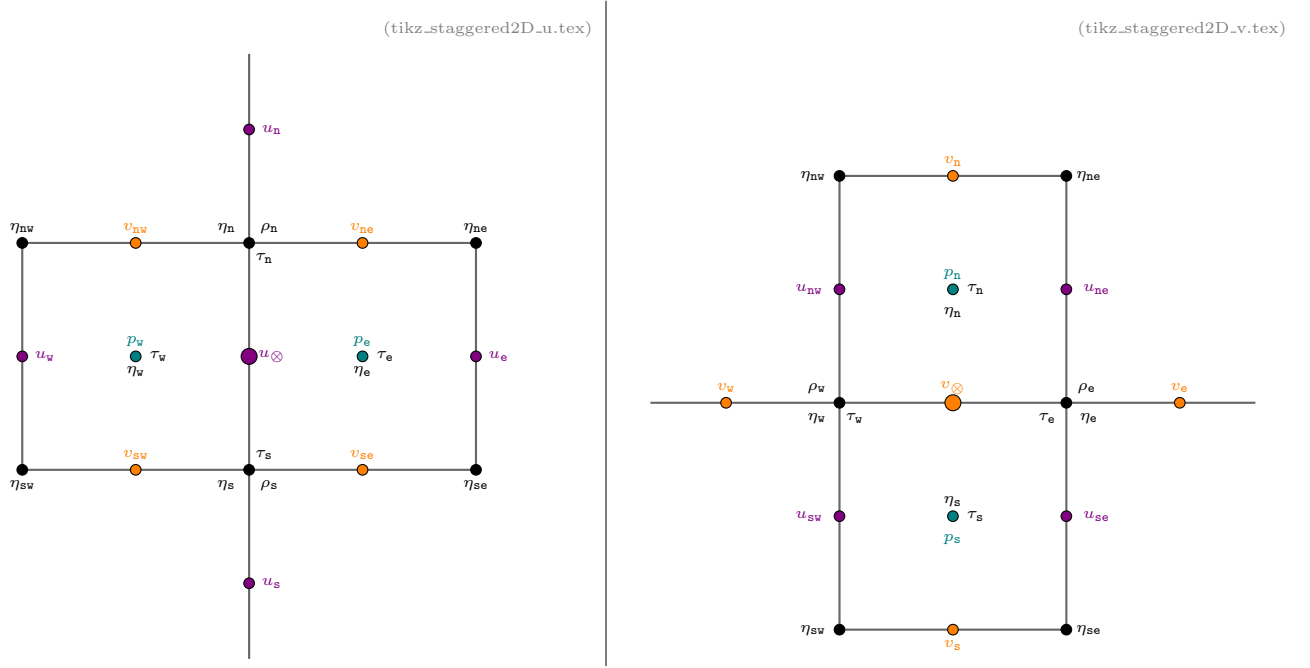
On the left side, we see that it would be trivial to compute $\partial u/\partial x$ and $\partial u/\partial y$ on node $u_\otimes$:

$$\frac{\partial u}{\partial x}\Big|_\otimes \quad \simeq \quad \frac{u_e - u_w}{2h_x}$$

$$\frac{\partial u}{\partial y}\Big|_\otimes \quad \simeq \quad \frac{u_n - u_s}{2h_y}$$

Likewise, on the right we could easily compute

$$\frac{\partial v}{\partial x}\Big|_\otimes \quad \simeq \quad \frac{v_e - v_w}{2h_x}$$

$$\frac{\partial v}{\partial y}\Big|_\otimes \quad \simeq \quad \frac{v_n - v_s}{2h_y}$$

Conversely, on the left side it would be difficult to compute any v derivative because no horizontal or vertical line passing by the u node in the middle intersects with a v node, and vice versa on the right side.

It then looks like our best option to compute $\dot\varepsilon_{xy}$ terms is actually on the background (black) nodes since on the left we would have

$$\dot\varepsilon_{xy}\Big|_\mathrm{n} = \frac{1}{2}\left(\frac{\partial u}{\partial y}\Big|_\mathrm{n} + \frac{\partial v}{\partial x}\Big|_\mathrm{n} \right) \simeq \frac{1}{2}\left(\frac{u_\mathrm{n} - u_\otimes}{h_y} + \frac{v_\mathrm{ne} - v_\mathrm{nw}}{h_x} \right)$$

This is indeed what we find in Gerya's book at p. 137:

229
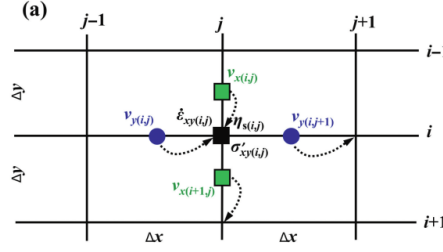
Taken from Gerya (2019).

However, we need to be careful with those nodes on the boundary, and even more careful with those at the corners of the domain.

Let us look again at our prototype mesh:
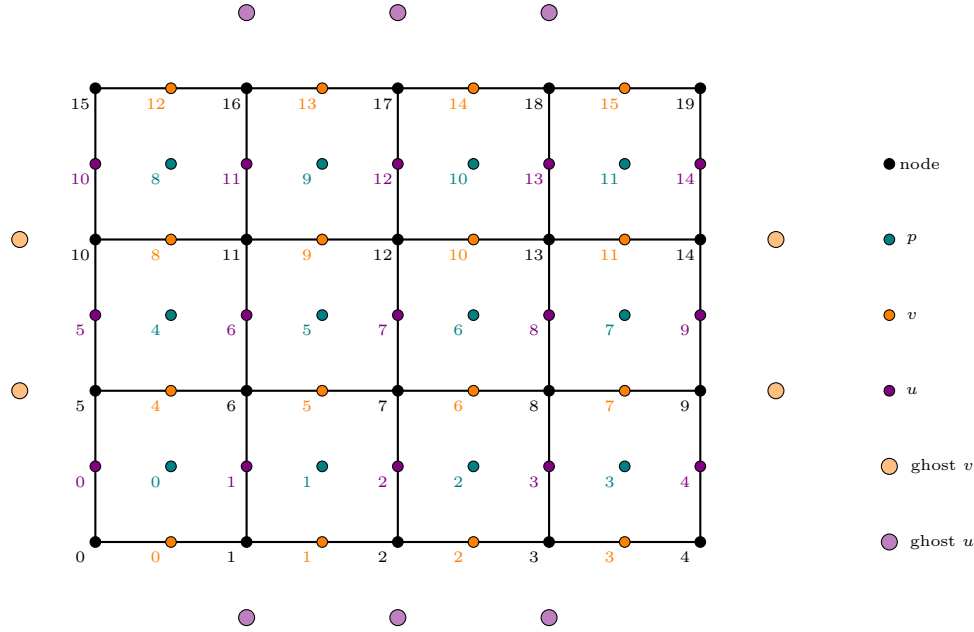
Let us consider node 10. It is 'missing' a left neighbour, which is the ghost node of v_8. Still operating under the assumption that boundary conditions are either free slip or no slip, the velocity assigned to this ghost is $\delta_{bc}v_8$, so that

$$\dot{\varepsilon}_{xy}|_{10} = \frac{1}{2}\left(\frac{\partial u}{\partial y}|_{10} + \frac{\partial v}{\partial x}|_{10}\right) \simeq \frac{1}{2}\left(\frac{u_{10} - u_5}{h_y} + \frac{v_8 - \delta_{bc}u_8}{h_x}\right)$$

Likewise, for node 3, we see that the ghost node of v_3 will be needed:

$$\dot{\varepsilon}_{xy}|_3 = \frac{1}{2}\left(\frac{\partial u}{\partial y}|_3 + \frac{\partial v}{\partial x}|_3\right) \simeq \frac{1}{2}\left(\frac{u_3 - \delta_{bc}u_3}{h_y} + \frac{v_3 - u_2}{h_x}\right)$$

As for the corners, let us consider node 15. We would need the ghost node of v_{12}, but since it would be on the top boundary its v value would also be zero. Likewise we would need the ghost node of u_{10} but it would be on the left boundary so its u value would be zero. In the end we find that $\dot{\varepsilon}_{xy} = 0$ in all four corners.

6.12.13 Embedding of Particle-In-Cell

how to interpolate fields onto particles?

section 8.4, 8.5 of book

6.12.14 About measuring discretisation errors

This subsection is borrowed from Duretz, May, Gerya, and Tackley [375].

We here focus on measuring the order of convergence of the primitive variables of the discrete Stokes problem.

In order to compute the discretization error, we utilize the L_1 norm, which for a scalar quantity ϕ, is defined as follows:

$$\|\phi\|_1 = \int_\Omega |\phi| \, dV$$

where Ω is the model domain. For a vector quantity $\vec{v} = (u, v)$ we have

$$\|\vec{v}\|_1 = \|u\|_1 + \|v\|_1 = \int_\Omega (|u| + |v|) \, dV$$

The pressure error is then defined as

$$\|e_p\|_1 = \|p - p_{mms}\|_1$$

where p_{mms} is the pressure of the manufactured solution.

The L_1 error for the velocity $\vec{v} = (u, v)$, is defined via

$$\|e_{\vec{v}}\|_1 = \|e_u\|_1 + \|e_v\|_1 = \|u - u_{mms}\|_1 + \|v - v_{mms}\|_1$$

where u_{mms} and v_{mms} are the exact values of u and v velocity components.

For numerical computations we approximate the above integrals via a one point quadrature rule (see Section 7), i.e.,

$$\|\phi\|_1 \simeq \|\phi\|_1^h = \sum_e |\phi(x_e, y_e)| \, V_e$$

where V_e is the representative volume for the point (x_e, y_e) in the middle of each cell e.

For FD schemes, the appropriate volume to use in the equation above is the control volume associated with each node. Within staggered grid FD schemes, the control volume associated with the p and u, v degrees of freedom are different.

Looking at the 4×3 mesh above, the volume associated with node u_0 is only $h_x h_y/2$ while the volume associated with node u_1 is $h_x h_y$. Same for v nodes on the boundaries and inside the domain. Only pressure node *all* are associated the same volume $h_x h_y$.

We define the following discrete L_1 norm for pressure on an $N \times N$-cell mesh as

$$\|p\|_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} |p_{i,j}| h_x h_y$$

where $h_x h_y$ is the cell volume. For the velocity components we have:

$$\|u\|_1 = \sum_{j=1}^{N} |u_{1,j}| \frac{h_x h_y}{2} + \sum_{i=2}^{N} \sum_{j=1}^{N} |u_{i,j}| h_x h_y + \sum_{j=1}^{N} |u_{N+1,j}| \frac{h_x h_y}{2}) \tag{6.140}$$

$$\|v\|_1 = \sum_{i=1}^{N} |v_{i,1}| \frac{h_x h_y}{2} + \sum_{i=1}^{N} \sum_{j=2}^{N} |v_{i,j}| h_x h_y + \sum_{i=1}^{N} |v_{i,N+1}| \frac{h_x h_y}{2}) \tag{6.141}$$

Actually, let us consider yet again our 4×3 mesh. We then have

$$
\begin{aligned}
\|u\|_1 &= \frac{1}{2}\sum_{j=1}^{N}|u_{0,j}|h_x h_y + \sum_{j=1}^{N}|u_{2,j}|h_x h_y + \sum_{j=1}^{N}|u_{3,j}|h_x h_y + \sum_{j=1}^{N}|u_{4,j}|h_x h_y + \frac{1}{2}\sum_{j=1}^{N}|u_{5,j}|h_x h_y \\
&= \sum_{j=1}^{N}\left[\frac{1}{2}|u_{1,j}| + |u_{2,j}| + |u_{3,j}| + |u_{4,j}| + \frac{1}{2}|u_{5,j}|\right]h_x h_y \\
&= \sum_{j=1}^{N}\left[\underbrace{\frac{1}{2}|u_{1,j}| + \frac{1}{2}|u_{2,j}|}_{\tilde{u}_{1,j}} + \underbrace{\frac{1}{2}|u_{2,j}| + \frac{1}{2}|u_{3,j}|}_{\tilde{u}_{2,j}} + \underbrace{\frac{1}{2}|u_{3,j}| + \frac{1}{2}|u_{4,j}|}_{\tilde{u}_{3,j}} + \underbrace{\frac{1}{2}|u_{4,j}| + \frac{1}{2}|u_{5,j}|}_{\tilde{u}_{4,j}}\right]h_x h_y \\
&= \sum_{e}\tilde{u}_e V_e \qquad\qquad\qquad (6.142)
\end{aligned}
$$

where $\tilde{u}$ is the velocity computed in the middle of each cell as the average of the absolute u velocities on each vertical side of the cell. This formulation is then much more practical to implement than the one consisting of three sums (see Eq. (6.140)) The same procedure can be repeated for the vertical component of the velocity.

The order of convergence of the discretization is determined by computing the numerical solution defined on a sequence of uniformly refined grids, and computing the L_1 norm of e_u and e_p on each grid. Following this, a least squares regression is performed on the log10 of the error norm and the cell size h. We relate the convergence rate of the L_1 norms for velocity and pressure errors to the order of convergence via

$$
\|e_u\|_1 \leq C \cdot h^{r_u} \qquad \|e_p\|_1 \leq C \cdot h^{r_p}
$$

where h is the mesh size, C is a constant independent of the grid resolution h, and r_u, r_p are the order of convergence for velocity and pressure fields, respectively.

write smthg about computing vrms

6.12.15 Non-equidistant nodes

Although less flexible than the FE method, the FDM allows for cells that are not all the same size. This is not fundamentally more difficult than what we have seen above but one now has to be careful about the meaning of the h_x and h_y quantities when discretising the equations.
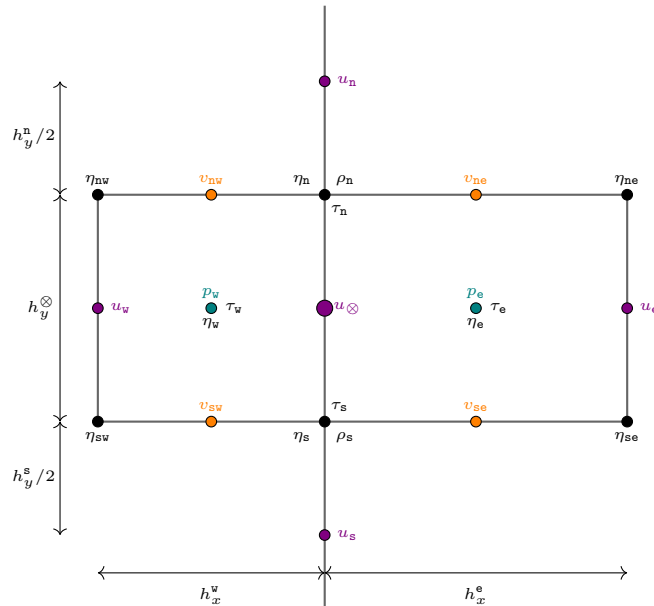
We start from

$$\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} - \frac{\partial p}{\partial x} + \rho g_x = 0$$

which we discretise as follows on the purple node in the middle:

$$\frac{\tau_{xx,\mathtt{e}} - \tau_{xx,\mathtt{w}}}{\frac{1}{2}h_x^{\mathtt{w}} + \frac{1}{2}h_x^{\mathtt{e}}} + \frac{\tau_{xy,\mathtt{n}} - \tau_{xy,\mathtt{s}}}{h_y^{\otimes}} - \frac{p_{\mathtt{e}} - p_{\mathtt{w}}}{\frac{1}{2}h_x^{\mathtt{w}} + \frac{1}{2}h_x^{\mathtt{e}}} = -\frac{\rho_{\mathtt{n}} + \rho_{\mathtt{s}}}{2}g_x$$

where

$$\tau_{xx,\mathtt{e}} = 2\eta_{\mathtt{e}}\dot{\varepsilon}_{xx,\mathtt{e}} = 2\eta_{\mathtt{e}}\frac{u_{\mathtt{e}} - u_{\otimes}}{h_x^{\mathtt{e}}}$$

$$\tau_{xx,\mathtt{w}} = 2\eta_{\mathtt{w}}\dot{\varepsilon}_{xx,\mathtt{w}} = 2\eta_{\mathtt{w}}\frac{u_{\otimes} - u_{\mathtt{w}}}{h_x^{\mathtt{w}}}$$

$$\tau_{xy,\mathtt{n}} = 2\eta_{\mathtt{n}}\dot{\varepsilon}_{xy,\mathtt{n}} = \eta_{\mathtt{n}}\left(\frac{u_{\mathtt{n}} - u_{\otimes}}{\frac{1}{2}h_y^{\mathtt{n}} + \frac{1}{2}h_y^{\otimes}} + \frac{v_{\mathtt{ne}} - v_{\mathtt{nw}}}{\frac{1}{2}h_x^{\mathtt{e}} + \frac{1}{2}h_x^{\mathtt{w}}}\right)$$

$$\tau_{xy,\mathtt{s}} = 2\eta_{\mathtt{s}}\dot{\varepsilon}_{xy,\mathtt{s}} = \eta_{\mathtt{s}}\left(\frac{u_{\otimes} - u_{\mathtt{s}}}{\frac{1}{2}h_y^{\otimes} + \frac{1}{2}h_y^{\mathtt{s}}} + \frac{v_{\mathtt{se}} - v_{\mathtt{sw}}}{\frac{1}{2}h_x^{\mathtt{e}} + \frac{1}{2}h_x^{\mathtt{w}}}\right)$$
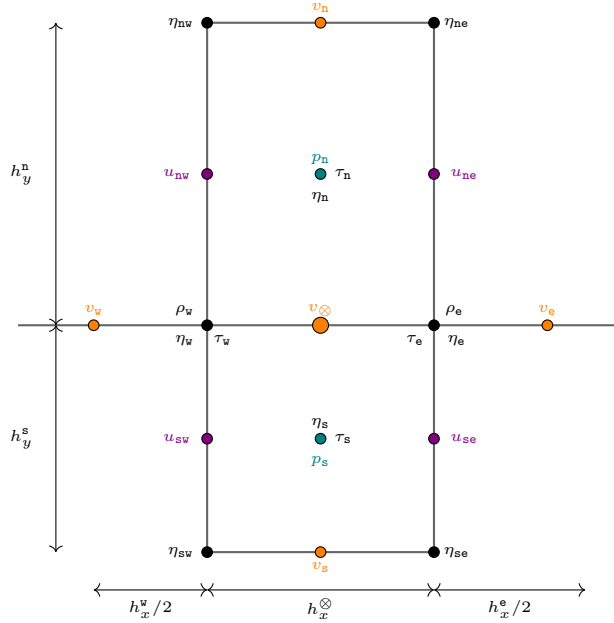
We start from

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} - \frac{\partial p}{\partial y} + \rho g_y = 0$$

which we discretise as follows on the $v_{\otimes}$ node in the middle

$$\frac{\tau_{xy,\mathtt{e}} - \tau_{xy,\mathtt{w}}}{h_x^{\otimes}} + \frac{\tau_{yy,\mathtt{n}} - \tau_{yy,\mathtt{s}}}{\frac{1}{2}h_y^{\mathtt{n}} + \frac{1}{2}h_y^{\mathtt{s}}} - \frac{p_{\mathtt{n}} - p_{\mathtt{s}}}{\frac{1}{2}h_y^{\mathtt{n}} + \frac{1}{2}h_y^{\mathtt{s}}} = -\frac{\rho_{\mathtt{e}} + \rho_{\mathtt{w}}}{2}g_y$$

where

$$
\begin{aligned}
\tau_{xy,\mathtt{e}} &= 2\eta_{\mathtt{e}}\dot{\varepsilon}_{xy,\mathtt{e}} = \eta_{\mathtt{e}}\left(\frac{u_{\mathtt{ne}} - u_{\mathtt{se}}}{\frac{1}{2}h_y^{\mathtt{n}} + \frac{1}{2}h_y^{\mathtt{s}}} + \frac{v_{\mathtt{e}} - v_{\otimes}}{\frac{1}{2}h_x^{\mathtt{e}} + \frac{1}{2}h_x^{\otimes}}\right) \\
\tau_{xy,\mathtt{w}} &= 2\eta_{\mathtt{w}}\dot{\varepsilon}_{xy,\mathtt{w}} = \eta_{\mathtt{w}}\left(\frac{u_{\mathtt{nw}} - u_{\mathtt{sw}}}{\frac{1}{2}h_y^{\mathtt{n}} + \frac{1}{2}h_y^{\mathtt{s}}} + \frac{v_{\otimes} - v_{\mathtt{w}}}{\frac{1}{2}h_x^{\otimes} + \frac{1}{2}h_x^{\mathtt{w}}}\right) \\
\tau_{yy,\mathtt{n}} &= 2\eta_{\mathtt{n}}\dot{\varepsilon}_{yy,\mathtt{n}} = 2\eta_{\mathtt{n}}\frac{v_{\mathtt{n}} - v_{\otimes}}{h_y^{\mathtt{n}}} \\
\tau_{yy,\mathtt{s}} &= 2\eta_{\mathtt{s}}\dot{\varepsilon}_{yy,\mathtt{s}} = 2\eta_{\mathtt{s}}\frac{v_{\otimes} - v_{\mathtt{s}}}{h_y^{\mathtt{s}}}
\end{aligned}
$$

What follows afterwards is not difficult: one needs to rederive the discretised mass and momentum conservation equations starting from the equations above.

6.13 Solving the wave equation

We start with the one-dimensional wave equation:

$$u_{tt} = c^2 u_{xx} \qquad x \in [0, L], \quad t \in [0, T]$$

which is supplemented by the initial conditions

$$
\begin{aligned}
u(x, 0) &= f(x) & \forall x \in [0, L] \\
u_t(x, 0) &= g(x) & \forall x \in [0, L]
\end{aligned}
\tag{6.143}
$$

and the boundary conditions

$$
\begin{aligned}
u(0, t) &= 0 & \forall t \in [0, T] \\
u_t(L, t) &= 0 & \forall t \in [0, T]
\end{aligned}
\tag{6.144}
$$

We then proceed to discretise the space domain with nnx equidistant nodes (forming $ncellx = nnx - 1$ cells) at locations x_i, $i = 0, ... nnx - 1$ and the time domain with N_t equidistant nodes at times t_i, $i = 1, ... N_t$. We define $h = L/ncellx$ and $\delta t = T/(N_t - 1)$

We are looking for the solution $u(x, t)$ at all mesh points in space in time and denote by u_i^k the discrete solution at location x_i and time t_k.

Following Section **??** the second-order derivatives will be replaced by central differences stencils, i.e. for a function f we would write

$$\frac{\partial^2 f}{\partial t^2} \simeq \frac{f^{k+1} - 2f^k + f^{k-1}}{\delta t^2}$$

6.13.1 Simple explicit method

The first idea is just to use central differences for both time and space derivatives, while assuming that the rhs (space) term is taken at time t_k: i.e.

$$\frac{u_i^{k+1} - 2u_i^k + u_i^{k-1}}{\delta t^2} = c^2 \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h^2}$$

We then define $\alpha = c\,\delta t/h$, so that we can write the equation above as

$$u_i^{k+1} - 2u_i^k + u_i^{k-1} = \alpha^2 (u_{i+1}^k - 2u_i^k + u_{i-1}^k) \tag{6.145}$$

or,

$$u_i^{k+1} = -u_i^{k-1} + 2(1 - \alpha^2)u_i^k + \alpha^2(u_{i+1}^k + u_{i-1}^k) \tag{6.146}$$

The parameter α is important and is often denoted by C or C_{CFL} as it is the Courant–Friedrichs–Lewy number[19].

Of course, we need to be careful about the initial conditions. Indeed, looking at the left hand side of Eq. (6.157), we see that we have three values of u_i at three consecutive different times. In other words, in order to compute u_i^{k+1} we need two previous values of u_i, i.e. u_i^{k-1} and u_i^k.

---

[19]https://en.wikipedia.org/wiki/Courant-Friedrichs-Lewy_condition

The relationship above can be written for any node inside the domain but not for those on the boundaries $(x = 0, L)$. However this is not a problem since we need not to write such an equation as the values of u at the extremities of the domain are given by the boundary conditions of Eq. (6.156).

The initial condition on u is trivial to implement. But quid of u_t? In this case we would write a centered difference as

$$g(x_i) = \left.\frac{\partial u}{\partial t}\right|_i^0 \simeq \frac{u_i^1 - u_i^{-1}}{2\delta t}$$

so that

$$u_i^{-1} = u_i^1 - 2\delta t \, g(x_i)$$

and then, when inserted in Eq. (6.146), we get for $k = 0$

$$
\begin{aligned}
u_i^1 &= -u_i^{-1} + 2(1 - \alpha^2)u_i^0 + \alpha^2(u_{i+1}^0 + u_{i-1}^0) \\
&= -(u_i^1 - 2\delta t g(x_i)) + 2(1 - \alpha^2)u_i^0 + \alpha^2(u_{i+1}^0 + u_{i-1}^0) \\
2u_i^1 &= 2\delta t g(x_i) + 2(1 - \alpha^2)u_i^0 + \alpha^2(u_{i+1}^0 + u_{i-1}^0) \\
u_i^1 &= \delta t g(x_i) + (1 - \alpha^2)f(x_i) + \frac{\alpha^2}{2}(f(x_{i+1}) + f(x_{i-1}))
\end{aligned}
\tag{6.147}
$$

We now have all discrete u values for the first and second time step so we can apply the algorithm delineated by Eq. (6.146) to arrive at u at subsequent time steps.

FIG 1 of Langtangen

explain that this is conditionally stable.

6.13.2   Implicit method(s)

We could overcome the problem of conditional stability by introducing an implicit scheme as follows (this time the space derivative terms are taken at time step $k + 1$):

$$\frac{u_i^{k+1} - 2u_i^k + u_i^{k-1}}{\delta t^2} = c^2 \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{h^2}$$

which leads to

$$u_i^{k+1} - 2u_i^k + u_i^{k-1} = \alpha^2(u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1})$$

or,

$$-\alpha^2 u_{i+1}^{k+1} + (1 + 2\alpha^2)u_i^{k+1} - \alpha^2 u_{i-1}^{k+1} = 2u_i^k - u_i^{k-1} \tag{6.148}$$

Another approach, the Crank-Nicolson method[20]:

$$\frac{u_i^{k+1} - 2u_i^k + u_i^{k-1}}{\delta t^2} = c^2 \frac{1}{2}\left(\frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{h^2} + \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h^2}\right)$$

$$u_i^{k+1} - 2u_i^k + u_i^{k-1} = \frac{\alpha^2}{2}\left(u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1} + u_{i+1}^k - 2u_i^k + u_{i-1}^k\right)$$

or,

$$(1 + \alpha^2)u_i^{k+1} - \frac{\alpha^2}{2}(u_{i+1}^{k+1} + u_{i-1}^{k+1}) = (2 - \alpha^2)u_i^k + \frac{\alpha^2}{2}(u_{i+1}^k + u_{i-1}^k) - u_i^{k-1} \tag{6.149}$$

Note that we have already encountered this method in the context of the diffusion equation solved with FDM in Section 6.6.

---

[20] https://en.wikipedia.org/wiki/Crank-Nicolson_method

6.13.3 Reflecting boundaries

When a wave hits a boundary, either it encounters a (Dirichlet) boundary condition as seen so far, but it can also encounter a different type of boundary condition: it is reflected back via a Neumann boundary condition, which translates into:

$$\frac{\partial u}{\partial n} = 0, \qquad \text{or} \quad \vec{n} \cdot \vec{\nabla} u = 0$$

In the case of a 1D domain the normal is $-\vec{e}_x$ at $x = 0$ and $+\vec{e}_x$ at $x = L$, which then translates into

$$-\frac{\partial u}{\partial x}\bigg|_{x=0} = 0$$

$$\frac{\partial u}{\partial x}\bigg|_{x=L} = 0$$

Having established the nature of these boundaries, we must now look at their implementation. For accuracy reasons we would like to resort to a central difference scheme (see Section XXX). However, this poses a problem, as visible when focusing on the left boundary (the time here is not important):

$$-\frac{\partial u}{\partial x}\bigg|_{x=0} \simeq -\frac{u_1 - u_{-1}}{2h} = 0 \tag{6.150}$$

Obviously u_{-1} does not exist! What we can do is write Eq. (6.146) at node $i = 0$:

$$u_0^{k+1} = -u_0^{k-1} + 2(1 - \alpha^2)u_0^k + \alpha^2(u_1^k + u_{-1}^k)$$

and insert in it u_{-1} obtained from Eq. (6.150):

$$\begin{aligned} u_0^{k+1} &= -u_0^{k-1} + 2(1 - \alpha^2)u_0^k + \alpha^2(u_1^k + u_1^k) \\ &= -u_0^{k-1} + 2u_0^k + 2\alpha^2(u_1^k - 2u_0^k) \end{aligned} \tag{6.151}$$

Likewise, on the right boundary (i.e. at node $nnx - 1$), we have

$$\frac{\partial u}{\partial x}\bigg|_{x=L} \simeq \frac{u_{nnx} - u_{nnx-2}}{2h} = 0$$

which we now insert in Eq. (6.146) expressed at $i = nnx - 1$:

$$\begin{aligned} u_{nnx-1}^{k+1} &= -u_{nnx-1}^{k-1} + 2(1 - \alpha^2)u_{nnx-1}^k + \alpha^2(u_{nnx-2}^k + u_{nnx-2}^k) \\ &= -u_{nnx-1}^{k-1} + 2u_{nnx-1}^k + 2\alpha^2(u_{nnx-2}^k - u_{nnx-1}^k) \end{aligned} \tag{6.152}$$

In his lecture notes Langtangen proposes the following rather elegant impementation[21]:

```
for i in range(0, Nx+1):
    ip1 = i+1 if i < Nx else i-1
    im1 = i-1 if i > 0 else i+1
    u[i] = u_1[i] + C2*(u_1[im1] - 2*u_1[i] + u_1[ip1])
```

adapt to my notations!

---

[21]I have adapted it to suit my notations

6.13.4 Generalisation: variable wave velocity

In the case when the wave velocity is not a constant in space, e.g. there are two media next to each other, the Laplace term is actually of the form

$$\frac{\partial}{\partial x}\left(c(x)^2 \frac{\partial u}{\partial x}\right)$$

which then needs to be discretised. The generic wave equation is then

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(c(x)^2 \frac{\partial u}{\partial x}\right) + h(x,t)$$

At a given node in time-space it writes

$$\frac{\partial^2}{\partial t^2} u(x_i, t_n) = \frac{\partial}{\partial x}\left(c(x_i)^2 \frac{\partial}{\partial x} u(x_i, t_n)\right) + h(x_i, t_n)$$

The typical way to approach this is to first define

$$\phi(x,t) = c(x)^2 \frac{\partial u}{\partial x}$$

so that a centered spatial derivative of ϕ writes

$$\frac{\partial \phi}{\partial x}(x_i, t_n) \simeq \frac{\phi(x_{i+1/2}, t_n) - \phi(x_{i-1/2}, t_n)}{h_x}$$

Of course we then turn our attention to (let us call $\underline{c} = c^2$ to simplify notations):

$$
\begin{aligned}
\phi(x_{i+1/2}, t_n) &= \underline{c}(x_{i+1/2}) \frac{\partial}{\partial x} u(x_{i+1/2}, t_n) \simeq \underline{c}(x_{i+1/2})^2 \frac{u_{i+1} - u_i}{h_x} \\
\phi(x_{i-1/2}, t_n) &= \underline{c}(x_{i-1/2}) \frac{\partial}{\partial x} u(x_{i-1/2}, t_n) \simeq \underline{c}(x_{i-1/2})^2 \frac{u_i - u_{i-1}}{h_x}
\end{aligned}
\tag{6.153}
$$

Putting it all together:

$$
\begin{aligned}
\frac{\partial \phi}{\partial x}(x_i, t_n) &\simeq \frac{1}{h_x}\left(\underline{c}(x_{i+1/2}) \frac{u_{i+1} - u_i}{h_x} - \underline{c}(x_{i-1/2}) \frac{u_i - u_{i-1}}{h_x}\right) \\
&\simeq \frac{1}{h_x^2}\left(\underline{c}(x_{i+1/2})(u_{i+1} - u_i) - \underline{c}(x_{i-1/2})(u_i - u_{i-1})\right)
\end{aligned}
\tag{6.154}
$$

But how can we compute $\underline{c}(x_{i+1/2})$? Three methods are commonly used.

- arithmetic averaging: $\underline{c}(x_{i+1/2}) \simeq \frac{1}{2}(\underline{c}(x_i) + \underline{c}(x_{i+1}))$

- geometric averaging: $\underline{c}(x_{i+1/2}) \simeq (\underline{c}(x_i)\underline{c}(x_{i+1}))^{1/2}$

- harmonic averaging: $\underline{c}(x_{i+1/2}) \simeq 2(1/\underline{c}(x_i) + 1/\underline{c}(x_{i+1}))^{-1}$

Of course similar expressions can then be used to compute $\underline{c}(x_{i-1/2})$.

6.13.5 Generalisation: damping

The wave equation that we have considered so far is rather ideal, in the sense that it is a bit too perfect: in reality waves energy is dissipated via a range of mechanims (think for instance at the air resistance in the case of a rope).

The simplest way to introduce damping is to add a first-order time derivative

$$\frac{\partial^2 u}{\partial t^2} + \gamma \frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} + h(x,t)$$

Obviously if one would remove the second-order time derivative we would recover a diffusion equation, which we have already studied in Section 6.4.

6.13.6   Periodic boundary conditions

6.13.7   (Von Neumann) Stability analysis

6.13.8   Numerical dispersion relation

6.13.9   Generalisation to 2D

We start with the two-dimensional wave equation:

$$u_{tt} = c^2(u_{xx} + u_{yy}) \qquad x \in \Omega, \quad t \in [0, T]$$

on the domain $\Omega = [0, L_x] \times [0, L_y]$ with boundary Γ, which is supplemented by the initial conditions

$$
\begin{aligned}
u(x, y, 0) &= f(x, y) & \forall (x, y) \in \Omega \\
u_t(x, y, 0) &= g(x, y) & \forall (x, y) \in \Omega
\end{aligned}
\tag{6.155}
$$

and the boundary conditions

$$
\begin{aligned}
u(x, y, t) &= 0 & \forall (x, y) \in \Gamma & \quad \forall t \in [0, T] \\
u_t(x, y, t) &= 0 & \forall (x, y) \in \Gamma & \quad \forall t \in [0, T]
\end{aligned}
\tag{6.156}
$$

We then proceed to discretise the space domain with nnx equidistant nodes (forming $ncellx = nnx - 1$ cells) in the x-direction at locations x_i, $i = 0, ...nnx - 1$, and with nny equidistant nodes (forming $ncelly = nny - 1$ cells) in the y-direction at locations y_j, $j = 0, ...nny - 1$, and the time domain with N_t equidistant nodes at times t_i, $i = 1, ...N_t$. We define $h_x = L_x/ncellx$, $h_y = L_y/ncelly$ and $\delta t = T/(N_t - 1)$.

Explicit formulation   The first idea is just to use central differences for both time and space derivatives, while assuming that the rhs (space) term is taken at time t_k: i.e.

$$\frac{u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1}}{\delta t^2} = c^2 \left(\frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{h_x^2} + \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k}{h_y^2} \right)$$

We then define $\alpha_x = c\,\delta t/h_x$, $\alpha_y = c\,\delta t/h_y$, so that we can write the equation above as

$$u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1} = \alpha_x^2(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k) + \alpha_y^2(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k) \tag{6.157}$$

$$\boxed{u_{i,j}^{k+1} = \alpha_x^2(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k) + \alpha_y^2(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k) + 2u_{i,j}^k - u_{i,j}^{k-1}}$$

Implicit formulations

$$\frac{u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1}}{\delta t^2} = c^2 \left(\frac{u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}}{h_x^2} + \frac{u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}}{h_y^2} \right)$$

$$u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1} = \alpha_x^2(u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}) + \alpha_y^2(u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1})$$

$$u_{i,j}^{k+1} - \alpha_x^2(u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}) - \alpha_y^2(u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}) = 2u_{i,j}^k - u_{i,j}^{k-1}$$

$$(1 + 2\alpha_x^2 + 2\alpha_y^2)u_{i,j}^{k+1} - \alpha_x^2(u_{i+1,j}^{k+1} + u_{i-1,j}^{k+1}) - \alpha_y^2(u_{i,j+1}^{k+1} + u_{i,j-1}^{k+1}) = 2u_{i,j}^k - u_{i,j}^{k-1}$$

Crank-Nicolson .

$$\frac{u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1}}{\delta t^2} = \frac{c^2}{2}\left(\frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{h_x^2} + \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k}{h_y^2}\right)$$
$$+ \frac{c^2}{2}\left(\frac{u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}}{h_x^2} + \frac{u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}}{h_y^2}\right)$$
$$u_{i,j}^{k+1} - 2u_{i,j}^k + u_{i,j}^{k-1} = \frac{\alpha_x^2}{2}\left(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k + u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}\right)$$
$$+ \frac{\alpha_y^2}{2}\left(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k + u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}\right)$$

$$u_{i,j}^{k+1} - \frac{\alpha_x^2}{2}\left(u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}\right) - \frac{\alpha_y^2}{2}\left(u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}\right)$$
$$= \frac{\alpha_x^2}{2}\left(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k\right) + \frac{\alpha_y^2}{2}\left(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k\right) + 2u_{i,j}^k - u_{i,j}^{k-1} \qquad (6.158)$$

$$(1 + \alpha_x^2 + \alpha_y^2)u_{i,j}^{k+1} - \frac{\alpha_x^2}{2}\left(u_{i+1,j}^{k+1} + u_{i-1,j}^{k+1}\right) - \frac{\alpha_y^2}{2}\left(u_{i,j+1}^{k+1} + u_{i,j-1}^{k+1}\right)$$
$$= \frac{\alpha_x^2}{2}\left(u_{i+1,j}^k + u_{i-1,j}^k\right) + \frac{\alpha_y^2}{2}\left(u_{i,j+1}^k + u_{i,j-1}^k\right) + (2 - \alpha_x^2 - \alpha_y^2)u_{i,j}^k - u_{i,j}^{k-1} \qquad (6.159)$$

In the case $h_x = h_y = h$ then $\alpha_x = \alpha_y = \alpha$ and we obtain

$$(1 + 2\alpha^2)u_{i,j}^{k+1} - \frac{\alpha^2}{2}\left(u_{i+1,j}^{k+1} + u_{i-1,j}^{k+1} + u_{i,j+1}^{k+1} + u_{i,j-1}^{k+1}\right)$$
$$= \frac{\alpha^2}{2}\left(u_{i+1,j}^k + u_{i-1,j}^k + u_{i,j+1}^k + u_{i,j-1}^k\right) + (2 - 2\alpha^2)u_{i,j}^k - u_{i,j}^{k-1} \qquad (6.160)$$

6.13.10   (von Neumann) Stability of the Crank-Nicolson method

This section is inspired by Degefa [344] (2018).

A numerical method is said to be stable if the cumulative effect of all the errors is bounded independently from the number of mesh points.

In order to establish whether a method is stable or not one often employs the so-called Von Neumann[22] stability analysis.

In general, the Von Neumann's procedure introduces an error represented by a finite Fourier series and examines how this error propagates during the solution.

To get stability of the 2D Crank-Nicolson method via this method we substitute $u_{i,j}$ in Eq. (6.160) by $u_{i,j} = \rho^k \exp(i\beta mh) \exp(i\gamma nh)$ in the homogeneous equation with h the mesh size in both directions, m, n spatial indices and β, γ time indices:

---

[22]https://en.wikipedia.org/wiki/John_von_Neumann

$$(1 + 2\alpha^2)u_{i,j}^{k+1} - \frac{\alpha^2}{2}\left(u_{i+1,j}^{k+1} + u_{i-1,j}^{k+1} + u_{i,j+1}^{k+1} + u_{i,j-1}^{k+1}\right)$$

$$= \frac{\alpha^2}{2}\left(u_{i+1,j}^{k} + u_{i-1,j}^{k} + u_{i,j+1}^{k} + u_{i,j-1}^{k}\right) + (2 - 2\alpha^2)u_{i,j}^{k} - u_{i,j}^{k-1} \qquad (6.161)$$

$$(1+2\alpha^2)\rho^{k+1}e^{i\beta mh}e^{i\gamma nh} - \frac{\alpha^2}{2}\left(\rho^{k+1}e^{i\beta(m+1)h}e^{i\gamma nh} + \rho^{k+1}e^{i\beta(m-1)h}e^{i\gamma nh} + \rho^{k+1}e^{i\beta mh}e^{i\gamma(n+1)h} + \rho^{k+1}e^{i\beta mh}e^{i\gamma(}\right.$$

$$= \frac{\alpha^2}{2}\left(\rho^{k}e^{i\beta(m+1)h}e^{i\gamma nh} + \rho^{k}e^{i\beta(m-1)h}e^{i\gamma nh} + \rho^{k}e^{i\beta mh}e^{i\gamma(n+1)h} + \rho^{k}e^{i\beta mh}e^{i\gamma(n-1)h}\right)$$

$$+(2-2\alpha^2)\rho^{k}e^{i\beta mh}e^{i\gamma nh} - \rho^{k-1}e^{i\beta mh}e^{i\gamma nh}$$

Dividing by $\rho^k e^{i\beta mh}e^{i\gamma nh}$ we obtain

$$(1+2\alpha^2)\rho - \frac{\alpha^2}{2}\left(\rho e^{i\beta h} + \rho e^{-i\beta h} + \rho e^{i\gamma h} + \rho e^{-i\gamma h}\right) = \frac{\alpha^2}{2}\left(e^{i\beta h} + e^{-i\beta h} + e^{i\gamma h} + e^{-i\gamma h}\right) + (2-2\alpha^2) - \rho^{-1}$$

We then recall that

$$\cos\theta = 1 - 2\sin^2\frac{\theta}{2} = \frac{e^{i\theta} + e^{-i\theta}}{2}$$

so that the equation above becomes

$$(1+2\alpha^2)\rho - \frac{\alpha^2}{2}\rho\left(\underbrace{e^{i\beta h} + e^{-i\beta h}}_{2\cos(\beta h)} + \underbrace{e^{i\gamma h} + e^{-i\gamma h}}_{2\cos(\gamma h)}\right) = \frac{\alpha^2}{2}\left(\underbrace{e^{i\beta h} + e^{-i\beta h}}_{2\cos(\beta h)} + \underbrace{e^{i\gamma h} + e^{-i\gamma h}}_{2\cos(\gamma h)}\right) + (2-2\alpha^2) - \rho^{-1}$$

$$(1+2\alpha^2)\rho - \alpha^2\rho\left[\cos(\beta h) + \cos(\gamma h)\right] = \alpha^2\left[\cos(\beta h) + \cos(\gamma h)\right] + (2-2\alpha^2) - \rho^{-1}$$

$$(1+2\alpha^2)\rho^2 - \alpha^2\rho^2\left[\cos(\beta h) + \cos(\gamma h)\right] - \alpha^2\rho\left[\cos(\beta h) + \cos(\gamma h)\right] - (2-2\alpha^2)\rho + 1 = 0$$

$$\rho^2\left[1 + \alpha^2(2 - \cos(\beta h) + \cos(\gamma h))\right] + \rho\left[-2 + \alpha^2(2 - \cos(\beta h) - \cos(\gamma h))\right] + 1 = 0$$

Using now the formula

$$1 - \cos\theta = 2\sin^2\frac{\theta}{2}$$

$$\rho^2\left[1 + \alpha^2\left(2\sin^2\frac{\beta h}{2}) + 2\sin^2\frac{\gamma h}{2}\right)\right] + \rho\left[-2 + \alpha^2\left(2\sin^2\frac{\beta h}{2}) + 2\sin^2\frac{\gamma h}{2}\right)\right] + 1 = 0$$

$$\rho^2\left[1 + 2\alpha^2\left(\sin^2\frac{\beta h}{2}) + \sin^2\frac{\gamma h}{2}\right)\right] + \rho\left[-2 + 2\alpha^2\left(\sin^2\frac{\beta h}{2}) + \sin^2\frac{\gamma h}{2}\right)\right] + 1 = 0$$

I do not understand the rest of the notes!

Chapter 7

Numerical integration

As we will see later, using the Finite Element method to solve problems involves computing integrals which are more often than not too complex to be computed analytically/exactly. We will then need to compute them numerically.

[wiki] In essence, the basic problem in numerical integration is to compute an approximate solution to a definite integral

$$I = \int_a^b f(x)dx \tag{7.1}$$

to a given degree of accuracy. This problem has been widely studied and we know that if $f(x)$ is a smooth function, and the domain of integration is bounded, there are many methods for approximating the integral to the desired precision.

There are several reasons for carrying out numerical integration.

- The integrand $f(x)$ may be known only at certain points, such as obtained by sampling. Some embedded systems and other computer applications may need numerical integration for this reason.

- A formula for the integrand may be known, but it may be difficult or impossible to find an antiderivative that is an elementary function. An example of such an integrand is $f(x) = \exp(-x^2)$, the antiderivative of which (the error function, times a constant) cannot be written in elementary form.

- It may be possible to find an antiderivative symbolically, but it may be easier to compute a numerical approximation than to compute the antiderivative. That may be the case if the antiderivative is given as an infinite series or product, or if its evaluation requires a special function that is not available.

Let us remember that the integral of Eq. (7.1) is in fact equal to the (signed) area between the x-axis and the curve $f(x)$ over the interval $[a, b]$:
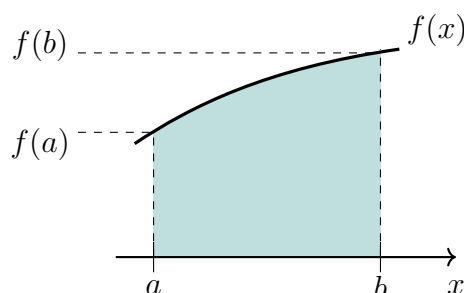
Note that in the example above $f(x) > 0$ so the area of the gray domain is counted positive. For example, if the function $f(x)$ is a polynomial the integral can easily be computed analytically. In the case of a 0th order polynomial, we have $f(x) = C$ where C is a constant. We then have

$$I = \int_a^b f(x)dx = \int_a^b C \, dx = C \int_a^b dx = C(b-a) \tag{7.2}$$
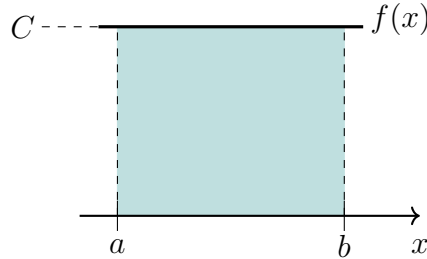
We see that the area of the gray domain in simply the product of its length $b - a$ by its height c and we indeed recover $I = C(b - a)$.

7.1 In 1 dimension

7.1.1 Midpoint and Trepezoidal rules

The simplest method of this type is to let the interpolating function be a constant function (a polynomial of degree zero) that passes through the point $((a+b)/2, f((a+b)/2))$. This is called the midpoint rule or rectangle rule. We then have

$$I = \int_a^b f(x)dx \simeq (b-a)f\left(\frac{a+b}{2}\right)$$

which is the area of this gray domain:
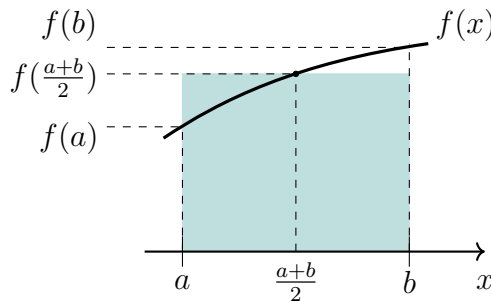
We can do a little bit better at virtually no cost: we choose the interpolating function to be a straight line (an affine function, i.e. a polynomial of degree 1) passing through the points $(a, f(a))$ and $(b, f(b))$. This is called the trapezoidal rule. Then

$$I = \int_a^b f(x)dx \simeq (b-a)\frac{f(a) + f(b)}{2}$$
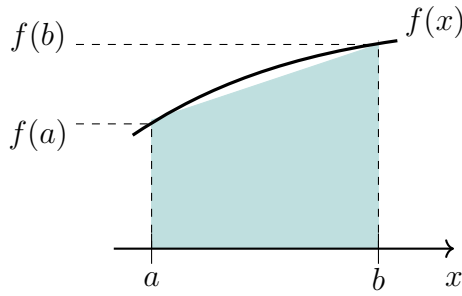
We see that if the function f is monotonous on the interval $[a, b]$ then the trapezoidal approach is likely to return a value close to the real value. However if the function f oscillates a lot in the interval, approximating it with a single rectangle or trapeze is not a sound assumption. We can then make use of the additive property of the integral: let c be the coordinate of the middle of the $[a, b]$ interval, i.e. $c = (a + b)/2$. Then we have

$$I = \int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

We can then apply the midpoint rule or the trapezoidal rule over both segments $[a, c]$ and $[c, b]$:
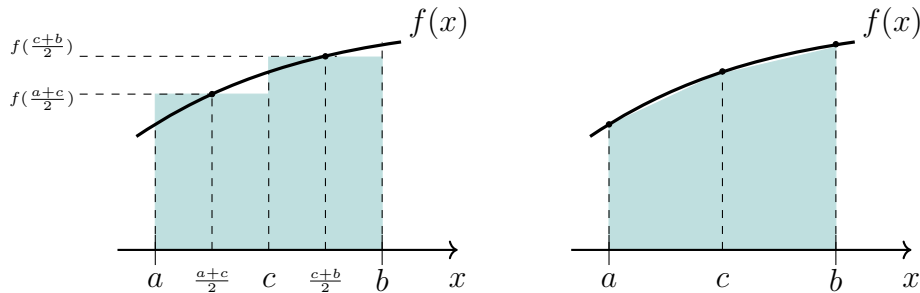
In this case we would have

$$
\begin{aligned}
I_{midpoint} &= (c - a)f(\frac{a + c}{2}) + (b - c)f(\frac{c + b}{2}) \\
I_{trapeze} &= (c - a)\frac{f(a) + f(c)}{2} + (b - c)\frac{f(c) + f(b)}{2}
\end{aligned}
$$

Of course we can repeat the process and for either one of these rules, we can make a more accurate approximation by breaking up the interval $[a, b]$ into some number n of subintervals, computing an approximation for each subinterval, then adding up all the results. For example, the composite trapezoidal rule can be stated as

$$\int_a^b f(x)dx \simeq \frac{b - a}{n}\left(\frac{f(a)}{2} + \sum_{k=1}^{n-1} f\left(a + k\frac{b - a}{n}\right) + \frac{f(b)}{2}\right) \tag{7.3}$$

where the subintervals have the form $[kh, (k + 1)h]$, with $h = (b - a)/n$ and $k = 0, 1, 2, \ldots, n - 1$.

add formula for mid-point and details
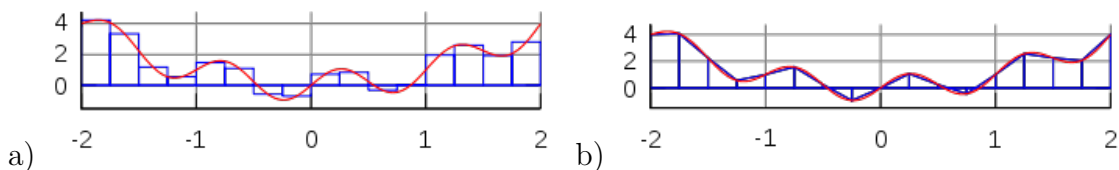
The interval $[-2, 2]$ is broken into 16 sub-intervals. The blue lines correspond to the approximation of the red curve by means of a) the midpoint rule, b) the trapezoidal rule.

There are several algorithms for numerical integration (also commonly called "numerical quadrature", or simply "quadrature") . Interpolation with polynomials evaluated at equally spaced points in $[a, b]$ yields the Newton-Cotes formulas, of which the rectangle rule and the trapezoidal rule are examples.

7.1.2   in 1D - Gauss-Legendre quadrature

If we allow the intervals between interpolation points to vary, we find another group of quadrature formulas, such as the Gauss(ian) quadrature formulas. A Gaussian quadrature rule is typically more accurate than a Newton-Cotes rule, which requires the same number of function evaluations, if the integrand is smooth (i.e., if it is sufficiently differentiable).

An $n-$point Gaussian quadrature rule, named after Carl Friedrich Gauss, is a quadrature rule constructed to yield an exact result for polynomials of degree $2n - 1$ or less by a suitable choice of the points x_i and weights w_i for $i = 1, \ldots, n$.

The domain of integration for such a rule is conventionally taken as $[-1, 1]$, so the rule is stated as

$$\int_{-1}^{+1} f(x)dx = \sum_{i_q=1}^{n} w_{i_q} f(x_{i_q})$$

In this formula the x_{i_q} coordinate is the i-th root of the Legendre polynomial[1] $P_n(x)$.

It is important to note that a Gaussian quadrature will only produce good results if the function $f(x)$ is well approximated by a polynomial function within the range $[-1, 1]$. As a consequence, the method is not, for example, suitable for functions with singularities.

---

[1]https://en.wikipedia.org/wiki/Legendre_polynomials

n	x_{iq}	w_{iq}	x_{iq} (approx)	w_{iq} (approx)
1	0	2	0	2
2	$\pm\sqrt{1/3}$	1	$\pm 0.577\ 350\ 269\ 189\ 626$	1
3	0	8/9	0	0.888 888 888 888 888
	$\pm\sqrt{3/5}$	5/9	$\pm 0.774\ 596\ 669\ 241\ 483$	0.555 555 555 555 555
4	$\pm\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{6/5}}$	$\frac{18+\sqrt{30}}{36}$	$\pm 0.339\ 981\ 043\ 584\ 856$	0.652 145 154 862 546
	$\pm\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{6/5}}$	$\frac{18-\sqrt{30}}{36}$	$\pm 0.861\ 136\ 311\ 594\ 953$	0.347 854 845 137 454
5	0	128/225	0	0.568 888 888 888 889
	$\pm\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}$	$\frac{322+13\sqrt{70}}{900}$	$\pm 0.538\ 469\ 310\ 105\ 683$	0.478 628 670 499 366
	$\pm\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}$	$\frac{322-13\sqrt{70}}{900}$	$\pm 0.906\ 179\ 845\ 938\ 664$	0.236 926 885 056 189
6	?	?	$\pm 0.238\ 619\ 186\ 083\ 197$	0.467 913 934 572 691
			$\pm 0.661\ 209\ 386\ 466\ 265$	0.360 761 573 048 139
			$\pm 0.932\ 469\ 514\ 203\ 152$	0.171 324 492 379 170
7			$\pm 0.949\ 107\ 912\ 342\ 759$	0.129 484 966 168 870
			$\pm 0.741\ 531\ 185\ 599\ 394$	0.279 705 391 489 277
			$\pm 0.405\ 845\ 151\ 377\ 397$	0.381 830 050 505 119
			$0.000\ 000\ 000\ 000\ 000$	0.417 959 183 673 469
8			$\pm 0.960\ 289\ 856\ 497\ 536$	0.101 228 536 290 376
			$\pm 0.796\ 666\ 477\ 413\ 627$	0.222 381 034 453 374
			$\pm 0.525\ 532\ 409\ 916\ 329$	0.313 706 645 877 887
			$\pm 0.183\ 434\ 642\ 495\ 650$	0.362 683 783 378 362
9			$\pm 0.968\ 160\ 239\ 507\ 626$	0.081 274 388 361 574
			$\pm 0.836\ 031\ 107\ 326\ 636$	0.180 648 160 694 857
			$\pm 0.613\ 371\ 432\ 700\ 590$	0.260 610 696 402 935
			$\pm 0.324\ 253\ 423\ 403\ 809$	0.312 347 077 040 003
			$0.000\ 000\ 000\ 000\ 000$	0.330 239 355 001 260
10			$\pm 0.973\ 906\ 528\ 517\ 172$	0.066 671 344 308 688
			$\pm 0.865\ 063\ 366\ 688\ 985$	0.149 451 349 150 581
			$\pm 0.679\ 409\ 568\ 299\ 024$	0.219 086 362 515 982
			$\pm 0.433\ 395\ 394\ 129\ 247$	0.269 266 719 309 996
			$\pm 0.148\ 874\ 338\ 981\ 631$	0.295 524 224 714 753

Abscissae and weights for Gauss quadratures up to $n = 10$. See [825, p89]. Also check `https://pomax.github.io/bezierinfo/legendre-gauss.html`.

As shown in the above table, it can be shown that the weight values must fulfil the following condition:

$$\sum_{i_q} w_{i_q} = 2 \tag{7.4}$$

This simply comes from the requirement that when $f(x) = 1$ then $\int_{-1}^{+1} f(x)dx = 2 = \sum w_{iq}$. It is also worth noting that all quadrature point coordinates are symmetrical around the origin.

Since most quadrature formula are only valid on a specific interval, we now must address the problem of their use outside of such intervals. The solution turns out to be quite simple: one must carry out a change of variables from the interval $[a, b]$ to $[-1, 1]$. We then consider the reduced coordinate $r \in [-1, 1]$ such that

$$r = \frac{2}{b-a}(x-a) - 1 \tag{7.5}$$

This relationship can be reversed such that when r is known, its equivalent coordinate $x \in [a, b]$ can

be computed:

$$x = \frac{b-a}{2}(1+r) + a \tag{7.6}$$

From this it follows that

$$dx = \frac{b-a}{2}dr \tag{7.7}$$

and then

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{+1} f(r)dr \simeq \frac{b-a}{2}\sum_{i_q=1}^{n_q} w_{i_q}f(r_{i_q}) \tag{7.8}$$

7.1.3   A probably naive way of finding the quadrature points coordinates and weights

We start from the assumption that the quadrature must be exact for polynomials $f(r)$, that it is written

$$I = \int_{-1}^{+1} f(r)dr = \sum_{i_q=1}^{n_q} w_{iq}f(r_{i_q})$$

and that $n_q > 0$, $w_{i_q} \neq 0$ and $r_{i_q} \in [-1, 1]$.

Let us start with zero-th order polynomials, i.e. $f(r) = C$. Then $I = 2C$ and we must then have

$$2C = \sum_{i_q=1}^{n_q} w_{iq}f(r_{i_q}) = \sum_{i_q=1}^{n_q} w_{iq}C$$

which imposes

$$\boxed{\sum_{i_q=1}^{n_q} w_{iq} = 2 \qquad \forall n_q > 0} \tag{7.9}$$

As long as the sum of the weights is equal to 2, any n_q-point based quadrature can integrate exactly a zero-th order polynomial.

Let us move on with first-order polynomials. Since we have covered the constant term hereabove, we set $f(r) = ar$ where $a \neq 0$. We have $I = 0$ so

$$0 = \sum_{i_q=1}^{n_q} w_{iq}f(r_{i_q}) = a\sum_{i_q=1}^{n_q} w_{iq}r_{i_q} \qquad \Rightarrow \qquad \boxed{\sum_{i_q=1}^{n_q} w_{iq}r_{i_q} = 0 \qquad \forall n_q > 0} \tag{7.10}$$

In order to integrate exactly first-order polynomials an n_q-point based quadrature must fulfil Eqs.(7.9) and (7.10).

- If $n_q = 1$, then we automatically have $w_1 = 2$ and $w_1 r_1 = 0$, i.e. $r_1 = 0$.

- If $n_q = 2$, then $w_1 + w_2 = 2$ and $w_1 r_1 + w_2 r_2 = 0$. There are many solutions w_1, w_2, r_1, r_2 which can fulfil these two equations, so this is not enough to determine a unique set of coordinates and weights.

Let us now turn to second-order polynomials: as before, I choose $f(r) = ar^2$. We have $I = 2a/3$ and

$$\frac{2a}{3} = \sum_{i_q=1}^{n_q} w_{iq}f(r_{i_q}) = a\sum_{i_q=1}^{n_q} w_{iq}r_{i_q}^2 \qquad \boxed{\sum_{i_q=1}^{n_q} w_{iq}r_{i_q}^2 = \frac{2}{3} \qquad \forall n_q > 0} \tag{7.11}$$

- If $n_q = 1$, we know that $w_1 = 2$ and $r_1 = 0$. This means that 1-point quadrature cannot exactly integrate polynomials higher than 1.

- If $n_q = 2$, then $w_1 + w_2 = 2$, $w_1 r_1 + w_2 r_2 = 0$ and now $w_1 r_1^2 + w_2 r_2^2 = 2/3$. We have three equations but still four unknowns. At this stage, we can do a simple additional assumption: common sense would have us realise that there is no reason why the (in this case) 2 quadrature point coordinates should be both negative or both positive. In light thereof we require that quadrature point coordinates are symmetric with respect with the origin $r = 0$, i.e. $r_1 = -r_2$ in this case. This yields to write: $w_1 r_1 + w_2 r_2 = w_1 r_1 + w_2(-r_1) = r_1(w_1 - w_2) = 0$. If $r_1 = 0$ then $r_2 = 0$ too and we do not have a 2-point quadrature. It must then follows that $w_1 = w_2$. And finally $w_1 r_1^2 + w_2 r_2^2 = w_1 r_1^2 + w_1(-r_1)^2 = 2/3$, i.e. $r_1 = -1/\sqrt{3}$ and $r_2 = 1/\sqrt{3}$ since $r_1 < r_2$.

If we now turn to third-order polynomials, i.e. $f(r) = ar^3$, then $I = 0$ again. We then must have

$$\boxed{\sum_{i_q=1}^{n_q} w_{iq} r_{i_q}^3 = 0 \qquad \forall n_q > 0} \tag{7.12}$$

We see that the coordinates and weights obtained for a 2-point quadrature verify this equation, i.e. a 2-point quadrature can also exactly integrate a 3rd-order polynomial. However, it is equally easy to verify that the 2-point quadrature cannot exactly integrate a 4th-order polynomial since

$$I = \int_{-1}^{+1} r^4 dr = \frac{2}{5} \neq \sum_{i_q=1}^{2} w_{iq} r_{i_q}^4$$

A three-point quadrature will then be needed for those. Because of the symmetry, we know that the middle point will be at $r = 0$.

Remark. *This approach unfortunately does not shed any light on why the method is called Gauss-Legendre quadrature nor why the quadrature points are the zeros of the Legendre polynomials...*

7.1.4 Examples

Example 1

Since we know how to carry out any required change of variables, we choose for simplicity $a = -1$, $b = +1$. Let us take for example $f(r) = \pi$. Then we can compute the integral of this function over the interval $[a, b]$ exactly:

$$I = \int_{-1}^{+1} f(r) dr = \pi \int_{-1}^{+1} dr = 2\pi$$

We can now use a Gauss-Legendre formula to compute this same integral:

$$I_{gq} = \int_{-1}^{+1} f(r) dr = \sum_{i_q=1}^{n_q} w_{i_q} f(r_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q} \pi = \pi \underbrace{\sum_{i_q=1}^{n_q} w_{i_q}}_{=2} = 2\pi$$

where we have used the property of the weight values of Eq.(7.4). Since the actual number of points was never specified, this result is valid for all quadrature rules.

Example 2

Let us now take $f(r) = mr + p$ and repeat the same exercise:

$$I = \int_{-1}^{+1} f(r)dr = \int_{-1}^{+1} (mr + p)dr = [\frac{1}{2}mr^2 + pr]_{-1}^{+1} = 2p$$

$$I_{gq} = \int_{-1}^{+1} f(r)dr = \sum_{i_q=1}^{n_q} w_{i_q}f(r_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q}(mr_{i_q} + p) = m\underbrace{\sum_{i_q=1}^{n_q} w_{i_q}r_{i_q}}_{=0} + p\underbrace{\sum_{i_q=1}^{n_q} w_{i_q}}_{=2} = 2p$$

since the quadrature points are symmetric w.r.t. to zero on the r-axis. Once again the quadrature is able to compute the exact value of this integral: this makes sense since an n-point rule exactly integrates a $2n - 1$ order polynomial such that a 1 point quadrature exactly integrates a first order polynomial like the one above.

Example 3

Let us now take $f(r) = r^2$. We have

$$I = \int_{-1}^{+1} f(r)dr = \int_{-1}^{+1} r^2 dr = \left[\frac{1}{3}r^3\right]_{-1}^{+1} = \frac{2}{3}$$

and

$$I_{gq} = \int_{-1}^{+1} f(r)dr = \sum_{i_q=1}^{n_q} w_{i_q}f(r_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q}r_{i_q}^2$$

- $n_q = 1$: $r_{i_q}^{(1)} = 0$, $w_{i_q} = 2$. $I_{gq} = 0$
- $n_q = 2$: $r_q^{(1)} = -1/\sqrt{3}$, $r_q^{(2)} = 1/\sqrt{3}$, $w_q^{(1)} = w_q^{(2)} = 1$. $I_{gq} = \frac{2}{3}$
- It also works $\forall n_q > 2$!

7.2 In 2 & 3 dimensions

7.2.1 On the reference square

Let us now turn to a two-dimensional integral of the form

$$I = \int_{-1}^{+1} \int_{-1}^{+1} f(r, s) dr ds$$

where $f(r, s)$ is again assumed to be continuous over the domain. The equivalent Gaussian quadrature writes:

$$I_{gq} \simeq \sum_{i_q=1}^{n_q} \sum_{j_q}^{n_q} f(r_{i_q}, s_{j_q}) w_{i_q} w_{j_q}$$

Finally we have

$$I = \int_a^{+b} \int_c^{+d} f(r, s) dr ds \simeq \frac{b-a}{2} \frac{d-c}{2} \sum_{i_q=1}^{n_q} \sum_{j_q}^{n_q} f(r_{i_q}, s_{j_q}) w_{i_q} w_{j_q} \qquad (7.13)$$

7.2.2 On a generic quadrilateral

Let K be a quadrilateral element with straight boundary lines and with vertices arranged as follows:
IMAGE
We wish to evaluate

$$I = \iint_K f(x, y) dx dy$$

In order to do so we will first transform the element K to the reference square element and then apply the quadrature of the previous section. This transformation can be carried out by means of the Q_1 basis functions, see Section 8.3.1. We construct a linear mapping between the quadrilateral element K and the reference square element:

$$x(r, s) = \sum_{i=1}^{4} \mathcal{N}_i(r, s) x_i \qquad (7.14)$$

$$y(r, s) = \sum_{i=1}^{4} \mathcal{N}_i(r, s) y_i \qquad (7.15)$$

Then we have

$$I = \iint_K f(x, y) dx dy = \int_{-1}^{+1} \int_{-1}^{+1} f(x(r, s), y(r, s)) |\boldsymbol{J}(r, s)| dr ds$$

where $\boldsymbol{J}(r, s)$ is the Jacobian of the transformation defined by

$$\boldsymbol{J}(r, s) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \end{pmatrix}$$

Finally applying the Gaussian quadrature yields:

$$I = \iint_K f(x, y) dx dy \simeq \sum_{i_q=1}^{n_q} \sum_{j_q}^{n_q} f(x(r_{i_q}, s_{j_q}), y(r_{i_q}, s_{j_q})) |\boldsymbol{J}(r_{i_q}, s_{j_q})| w_{i_q} w_{j_q}$$

7.2.3 Exercises

Exercise Quad-1

Write a program which uses the midpoint rule to compute (subdivide the interval in n subintervals)

$$I = \int_0^{\pi/2} f(x) \, dx \qquad f(x) = x \qquad \text{and} \qquad f(x) = \cos(x)$$

For this you will need to declare and use a function which either returns x or $\cos(x)$. Compute and plot the (absolute) error between the measured I_n and the analytical value I as a function of the subinterval size h.

Bonus: same as before with $I = \int_{y=1}^{3} \int_{x=2}^{4} (x^2 y^3 + xy + 1) dx dy$.

Exercise Quad-2

Same exercise as above but with the trapezoidal rule. Which method is the most accurate?

The following Fortran program is an example of how Gauss quadrature can be implemented:

```fortran
program integration
implicit none
integer, parameter:: nq=2
real(8),dimension(nq),parameter:: xq=(/-1.d0/sqrt(3.d0),+1.d0/sqrt(3.d0)/)
real(8),dimension(nq),parameter:: wq=(/1.d0,1.d0/)
real(8) I
integer iq

I=0.d0
do iq=1,nq
   I=I+wq(iq)*fct(xq(iq))
end do

write(*,*) 'I=',I

contains

function fct(x)
implicit none
real(8) x,fct

!fct=3.14
!fct=x+1
fct=x**2

end function

end program
```

Modify/translate this previous program to use 5 quadrature points instead of two.

Integrate the functions

$$f_1(x) = \sin(x\pi + \pi/2) \qquad f_2(x) = \sqrt{x+1} \qquad f_3(x) = x^4 - x^3$$

with the 2-point and the 5-point quadrature rules.
Compare the results with the analytical values.

Compute analytically the integral of the function $f(x,y) = x^2 + 4y$ over the domain $\Omega = [11, 14] \times [7, 10]$.
Write a code which integrates this function by means of a 2×2, 3×3 or 4×4 Gauss-Legendre quadrature algorithm.

7.2.4 Quadrature on triangles

Our goal is to develop a quadrature rule of the form

$$\iint_\triangle f(r,s)\,drds \simeq \frac{1}{2}\sum_{i_q=1}^{n_q}\omega_{i_q}f(r_{i_q},s_{i_q})$$

We will here add two requirements: a) we would like to find quadrature rules which achieve the highest possible accuracy for the lowest possible number of quadrature points; b) we would like the quadrature points to possess some kind of symmetry. Note that the factor $1/2$ in the equation above is a convention. If $f=1$ then the left hand term is the area of the triangle which is $1/2$. Since people usually require that $\sum_i \omega_i = 1$ then the factor $1/2$ is necessary.

Before we go any further, we need to establish that

$$\iint_\triangle \{1,r,s,r^2,rs,t^2,r^3,r^2s,rs^2,s^3\}\,drds = \left\{\frac{1}{2},\frac{1}{6},\frac{1}{6},\frac{1}{12},\frac{1}{24},\frac{1}{12},\frac{1}{20},\frac{1}{60},\frac{1}{60},\frac{1}{20}\right\}$$

where $\triangle$ stands for the reference triangle.

Gaussian quadrature of degree 1

This means that the quadrature should be accurate for $f(r,s) = \{1,r,s\}$. We then obtain:

$$\iint_\triangle 1\,drds = \frac{1}{2} = \frac{1}{2}\sum_{i_q=1}^{1}\omega_{i_q}f(r_{i_q},s_{i_q}) = \frac{1}{2}\omega_1 f(r_1,s_1) = \frac{1}{2}\omega_1$$

$$\iint_\triangle r\,drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{1}\omega_{i_q}r_{i_q} = \frac{1}{2}\omega_1 r_1$$

$$\iint_\triangle s\,drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{1}\omega_{i_q}s_{i_q} = \frac{1}{2}\omega_1 s_1$$

We then obtain $\omega_1 = 1$ and $r_1 = s_1 = \frac{1}{3}$.

Gaussian quadrature of degree 2

The quadrature should be accurate for $f(r,s) = \{1, r, s, r^2, rs, t^2\}$ so that

$$\iint_\triangle 1\, drds = \frac{1}{2} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q}$$

$$\iint_\triangle r\, drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}$$

$$\iint_\triangle s\, drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} s_{i_q}$$

$$\iint_\triangle r^2\, drds = \frac{1}{12} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}^2$$

$$\iint_\triangle rs\, drds = \frac{1}{24} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q} s_{i_q}$$

$$\iint_\triangle s^2\, drds = \frac{1}{12} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} s_{i_q}^2$$

$$(7.16)$$

If we set $n_q = 1$ then we have 6 equations and only three unknowns ω_1, r_1, s_1 so this will not work. If we set $n_q = 2$ then we have 6 equations and six unknowns $\omega_1, r_1, s_1, \omega_2, r_2, s_2$ but we find that the quadrature is not symmetric[2]. If we set $n_q = 3$ then we have 6 equations and 9 unknowns and the solution is not unique. Because of symmetry we for instance have to impose $r_2 = s_3$ and $s_2 = r_3$ (points 2 and 3 are symmetric with respect to the $r = s$ line). Two common quadratures are found:

$$(r_1, s_1) = (\frac{1}{6}, \frac{1}{6}) \qquad (r_2, s_2) = (\frac{2}{3}, \frac{1}{6}) \qquad (r_3, s_3) = (\frac{1}{6}, \frac{2}{3}) \qquad \omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$$

and

$$(r_1, s_1) = (0, \frac{1}{2}) \qquad (r_2, s_2) = (\frac{1}{2}, 0) \qquad (r_3, s_3) = (\frac{1}{2}, \frac{1}{2}) \qquad \omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$$

[2]proof to do

Gaussian quadrature of degree 3

The quadrature should be accurate for $f(r,s) = \{1, r, s, r^2, rs, t^2, r^3, r^2s, rs^2, t^2\}$ so that

$$\iint_\triangle 1 drds = \frac{1}{2} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q}$$

$$\iint_\triangle r drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}$$

$$\iint_\triangle s drds = \frac{1}{6} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} s_{i_q}$$

$$\iint_\triangle r^2 drds = \frac{1}{12} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}^2$$

$$\iint_\triangle rs drds = \frac{1}{24} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q} s_{i_q}$$

$$\iint_\triangle s^2 drds = \frac{1}{12} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} s_{i_q}^2$$

$$\iint_\triangle r^3 drds = \frac{1}{20} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}^3$$

$$\iint_\triangle r^2 s drds = \frac{1}{60} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q}^2 s_{i_q}$$

$$\iint_\triangle rs^2 drds = \frac{1}{60} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} r_{i_q} s_{i_q}^2$$

$$\iint_\triangle s^3 drds = \frac{1}{20} = \frac{1}{2}\sum_{i_q=1}^{n_q} \omega_{i_q} s_{i_q}^3$$

$$(7.17)$$

This time $n_q = 3$ will not work since we have 10 equations. Switching to $n_q = 4$ we now have 12 unknowns and 10 equations so there are many possibilities. One of them is given in the table below. As a side note there could be cases where having a negative weight ω could be problematic.

Tabulated quadrature rules on the reference triangle

In what follows we use the following quadrature rule:

$$\iint_\triangle f(r,s)\, drds \simeq \sum_{i_q=1}^{n_q} \omega_{i_q} f(r_{i_q}, s_{i_q})$$

with $\sum_i \omega_i = 1/2$.

Quadrature rules for triangles can be found in Dunavant [373] (1985). The following ones are identical to those in the *ip_triangle.m* file of the MILAMIN code [318]. See also Lether [821] (1976) on the topic of computation of double integrals over a triangle.

	r_q exact	s_q exact	w_q exact	r_q approx.	s_q approx.	w_q approx.
$i_q=1$	$1/3$	$1/3$	$1/2$			
$i_q=1$	$1/6$	$1/6$	$1/6$			
$i_q=2$	$2/3$	$1/6$	$1/6$			
$i_q=3$	$1/6$	$2/3$	$1/6$			
$i_q=1$	$1/3$	$1/3$	$-27/96$			
$i_q=2$	0.6	0.2	$25/96$			
$i_q=3$	0.2	0.6	$25/96$			
$i_q=4$	0.2	0.2	$25/96$			
$i_q=1$	$1-2g_1$	g_1	$w_1/2$	0.108103018168070	0.44594849091596	
$i_q=2$	g_1	$1-2g_1$	$w_1/2$	0.445948490915965	0.108103018168070	
$i_q=3$	g_1	g_1	$w_1/2$	0.445948490915965	0.445948490915965	
$i_q=4$	$1-2g_2$	g_2	$w_2/2$	0.816847572980459	0.091576213509771	
$i_q=5$	g_2	$1-2g_2$	$w_2/2$	0.091576213509771	0.816847572980459	
$i_q=6$	g_2	g_2	$w_2/2$	0.091576213509771	0.091576213509771	
$i_q=1$				0.091576213509771	0.091576213509771	0.109951743655322/2.0
$i_q=2$				0.816847572980459	0.091576213509771	0.109951743655322/2.0
$i_q=3$				0.091576213509771	0.816847572980459	0.109951743655322/2.0
$i_q=4$				0.445948490915965	0.445948490915965	0.223381589678011/2.0
$i_q=5$				0.108103018168070	0.445948490915965	0.223381589678011/2.0
$i_q=6$				0.445948490915965	0.108103018168070	0.223381589678011/2.0
$i_q=1$				0.1012865073235	0.1012865073235	0.0629695902724
$i_q=2$				0.7974269853531	0.1012865073235	0.0629695902724
$i_q=3$				0.1012865073235	0.7974269853531	0.0629695902724
$i_q=4$				0.4701420641051	0.0597158717898	0.0661970763942
$i_q=5$				0.4701420641051	0.4701420641051	0.0661970763942
$i_q=6$				0.0597158717898	0.4701420641051	0.0661970763942
$i_q=7$				0.3333333333333	0.3333333333333	0.1125000000000
$i_q=1$				$5.01426509658179e-01$	$2.49286745170910e-01$	$5.83931378631895e-02$
$i_q=2$				$2.49286745170910e-01$	$5.01426509658179e-01$	$5.83931378631895e-02$
$i_q=3$				$2.49286745170910e-01$	$2.49286745170910e-01$	$5.83931378631895e-02$
$i_q=4$				$8.73821971016996e-01$	$6.30890144915020e-02$	$2.54224531851035e-02$
$i_q=5$				$6.30890144915020e-02$	$8.73821971016996e-01$	$2.54224531851035e-02$
$i_q=6$				$6.30890144915020e-02$	$6.30890144915020e-02$	$2.54224531851035e-02$
$i_q=7$				$5.31450498448170e-02$	$3.10352451033784e-01$	$4.14255378091870e-02$
$i_q=8$				$6.36502499121399e-01$	$5.31450498448170e-02$	$4.14255378091870e-02$
$i_q=9$				$3.10352451033784e-01$	$6.36502499121399e-01$	$4.14255378091870e-02$
$i_q=10$				$5.31450498448170e-02$	$6.36502499121399e-01$	$4.14255378091870e-02$
$i_q=11$				$6.36502499121399e-01$	$3.10352451033784e-01$	$4.14255378091870e-02$
$i_q=12$				$3.10352451033784e-01$	$5.31450498448170e-02$	$4.14255378091870e-02$

where

$$g_1 = \left(8 - \sqrt{10} + \sqrt{38 - 44\sqrt{2/5}}\right)/18 \qquad g_2 = \left(8 - \sqrt{10} - \sqrt{38 - 44\sqrt{2/5}}\right)/18$$

$$w_1 = \left(620 + \sqrt{213125 - 53320\sqrt{10}}\right)/3720 \qquad w_2 = \left(620 - \sqrt{213125 - 53320\sqrt{10}}\right)/3720$$

All these are implemented in STONE 120 (see also STONE 112).

The case of a generic triangle

Let T be a triangular element with straight edges defined by the three vertices (x_i, y_i) $(i = 1, 2, 3)$ arranged in the counter-clockwise order:

   INSERT FIGURE

   We now need to evaluate the following integral over T:

$$I = \iint_T f(x, y)\, dxdy$$

We will proceed similarly to the quadrilateral case: first transform ('map') this triangle into the reference triangle $\triangle$ and then use the adequate quadrature rule.

   We can base this transformation ('mapping') on the linear ('P_1') basis functions[3]:

$$
\begin{aligned}
\mathcal{N}_1(r, s) &= 1 - r - s \\
\mathcal{N}_1(r, s) &= r \\
\mathcal{N}_1(r, s) &= s
\end{aligned}
$$

with

$$x = P(r, s) = \sum_{i=1}^{3} x_i \mathcal{N}_i(r, s) = x_1 \mathcal{N}_1(r, s) + x_2 \mathcal{N}_2(r, s) + x_3 \mathcal{N}_3(r, s)$$

$$y = Q(r, s) = \sum_{i=1}^{3} y_i \mathcal{N}_i(r, s) = y_1 \mathcal{N}_1(r, s) + y_2 \mathcal{N}_2(r, s) + y_3 \mathcal{N}_3(r, s)$$

For example when $(r, s) \to (0, 0)$ we find $(x, y) \to (x_1, y_1)$, when $(r, s) \to (1, 0)$ we find $(x, y) \to (x_2, y_2)$, and when $(r, s) \to (0, 1)$ we find $(x, y) \to (x_3, y_3)$.

   Then we have

$$I = \iint_T f(x, y)\, dxdy = \iint_\triangle f(P(r, s), Q(r, s))\, |J(r, s)|\, drds$$

where $J(r, s)$ is the Jacobian of the transformation:

$$
J(r, s) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \end{vmatrix} = \begin{vmatrix} \frac{\partial}{\partial r}\sum_{i=1}^{3} x_i \mathcal{N}(r_i, s_i) & \frac{\partial}{\partial r}\sum_{i=1}^{3} y_i \mathcal{N}(r_i, s_i) \\ \frac{\partial}{\partial s}\sum_{i=1}^{3} x_i \mathcal{N}(r_i, s_i) & \frac{\partial}{\partial s}\sum_{i=1}^{3} y_i \mathcal{N}(r_i, s_i) \end{vmatrix} = \begin{vmatrix} (-x_1 + x_2) & (-y_1 + y_2) \\ (-x_1 + x_3) & (-y_1 + y_3) \end{vmatrix} = 2\mathcal{A}_T
$$

where $\mathcal{A}_T$ is the area of triangle T. In the end we have

$$I = \iint_T f(x, y)\, dxdy = 2\mathcal{A}_T \iint_\triangle f(P(r, s), Q(r, s))\, drds$$

and the rhs integral can then be computed by means of quadrature.

7.2.5   A mathematical recreation: computing the volume of a tetrahedron

Let us find the volume of tetrahedron bounded by the planes passing through the points A(1,0,0), B(0,1,0), C(0,0,1) and the coordinate planes Oxy, Oxz and Oyz.

---

[3]See Section 11
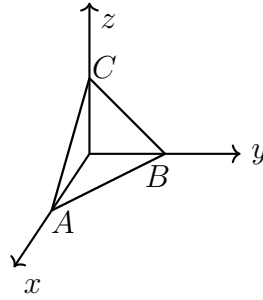
The equation of the plane is $x + y + z = 1$, or $z = 1 - x - y$. Hence, the limits of integration over the variable z range in the interval from $z = 0$ to $z = 1 - x - y$. Now we can calculate the volume of the tetrahedron:

$$
\begin{aligned}
V &= \iiint dx\, dy\, dz \\
&= \int_0^1 dx \int_0^{1-x} dy \int_0^{1-x-y} dz \\
&= \int_0^1 dx \int_0^{1-x} dy(1 - x - y) \\
&= \int_0^1 dx \left[y - xy - \frac{1}{2}y^2 \right]_0^{1-x} \\
&= \int_0^1 dx \left((1-x) - x(1-x) - \frac{1}{2}(1-x)^2 \right) \\
&= \int_0^1 dx \left(\frac{1}{2} - x + \frac{1}{2}x^2 \right) \\
&= \left[\frac{1}{2}x - \frac{1}{2}x^2 + \frac{1}{6}x^3 \right]_0^1 \\
&= \frac{1}{6}
\end{aligned}
\tag{7.18}
$$

We will use this result in the following section.

7.2.6   Quadrature on tetrahedra

Remark. *In what follows the coefficients in the tables are not the reduced coordinates of the quadratue points but the coefficients corresponding to the 4 nodes.*

Quadrature rules on tetrahedra take the form:

$$
\int\int\int_{el} f(x,y,z)dxdydz = V_{el} \sum_{iq=1}^{nqel} w_{iq} f(\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq})
$$

or, that is to say:

$$
\int\int\int_{el} f(x,y,z)dxdydz = \sum_{iq=1}^{nqel} (w_{iq} V_{el}) f(\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq})
$$

258

with in our case $V_{el} = 1/6$.

In the literature it can be found that a one point quadrature is characterised by

$$w_{iq} = 1 \qquad \xi_1^{iq} = \xi_2^{iq} = \xi_3^{iq} = \xi_4^{iq} = 0.25$$

i.e, the coordinates of the single point are given by:

$$x_{iq} = \sum_{i=1}^{4} \xi_i^{iq} x_i = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$$

Same for y and z coordinates.

A four-point quadrature rule is characterised by $w_{iq} = V_{el} * 0.25 = 1/24 \simeq 04166666666666667$ and

	ξ_1	ξ_2	ξ_3	ξ_4
iq=1	0.585410196624969	0.138196601125011	0.138196601125011	0.138196601125011
iq=2	0.138196601125011	0.585410196624969	0.138196601125011	0.138196601125011
iq=3	0.138196601125011	0.138196601125011	0.585410196624969	0.138196601125011
iq=4	0.138196601125011	0.138196601125011	0.138196601125011	0.585410196624969

We then have:

$$r_{iq} = \sum_{i=1}^{4} \xi_i^{iq} x_i = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (r_1, r_2, r_3, r_4) = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (0, 1, 0, 0) = \xi_2^{iq}$$

$$s_{iq} = \sum_{i=1}^{4} \xi_i^{iq} y_i = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (s_1, s_2, s_3, s_4) = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (0, 0, 1, 0) = \xi_3^{iq}$$

$$t_{iq} = \sum_{i=1}^{4} \xi_i^{iq} z_i = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (t_1, t_2, t_3, t_4) = (\xi_1^{iq}, \xi_2^{iq}, \xi_3^{iq}, \xi_4^{iq}) \cdot (0, 0, 0, 1) = \xi_4^{iq}$$

Finally:

	r_q	s_q	t_q	w_q
iq=1	0.138196601125011	0.138196601125011	0.138196601125011	0.04166666666666667
iq=2	0.585410196624969	0.138196601125011	0.138196601125011	0.04166666666666667
iq=3	0.138196601125011	0.585410196624969	0.138196601125011	0.04166666666666667
iq=4	0.138196601125011	0.138196601125011	0.585410196624969	0.04166666666666667

7.2.7   The Gauss-Lobatto approach

All what we have seen above falls under the Gauss-Legendre quadrature method. There is however another somewhat common quadrature method: the Gauss-Lobatto quadrature. . It is similar to Gaussian quadrature with the following important differences: 1) There are integration points in the interval but they also always include the end points of the integration interval; 2) It is accurate for polynomials up to degree $2n - 3$, where n is the number of integration points.

In 1D, it reads:

$$\int_{-1}^{+1} f(x)dx = \frac{2}{n(n-1)}[f(-1) + f(1)] + \sum_{i=2}^{n-1} w_i f(x_i)$$

The locations and weights of the integration points are as follows:

n	x_{iq}	w_{iq}	x_{iq} (approx)	w_{iq} (approx)
3	0	4/3		
	± 1	1/3		
4	$\pm\sqrt{\frac{1}{5}}$	5/6		
	± 1	1/6		
5	0	32/45		
	$\pm\sqrt{\frac{3}{7}}$	49/90		
	± 1	1/10		
6	$\pm\sqrt{\frac{1}{3} - \frac{2\sqrt{7}}{21}}$	$\frac{14+\sqrt{7}}{30}$		
	$\pm\sqrt{\frac{1}{3} + \frac{2\sqrt{7}}{21}}$	$\frac{14-\sqrt{7}}{30}$		
	± 1	1/15		

7.2.8 Computing the 'real' coordinates of the quadrature points and other considerations

The quadrature point coordinates are always given in (what I call) reduced coordinates, i.e. between -1 and 1. However, one sometimes need their equivalent in the x, y Cartesian space. This is trivial once one remembers that within an element, a field f is reprensented as follow:

$$f(r,s) = \sum_{i=1}^{m} \mathcal{N}_i(r,s) f_i$$

where m is the number of nodes, r and s are the reduced coordinates and $\mathcal{N}_i$ are the basis functions. The value of f at a quadrature point (r_q, s_q) is then simply

$$f(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q) f_i$$

If we now take $f = x$, then

$$x_q = x(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q) x_i$$

and

$$y_q = y(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q) y_i$$

where x_i and y_i are the Cartesian coordinates of the nodes. This is then easily extended to three dimensions:

$$x_q = x(r_q, s_q, t_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q, t_q) x_i$$

$$y_q = x(r_q, s_q, t_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q, t_q) y_i$$

$$z_q = z(r_q, s_q, t_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q, t_q) z_i$$

or,

$$\vec{r}_q = \vec{r}(r_q, s_q, t_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q, t_q) \vec{r}_i$$

This also applies to other fields such as velocity, temperature, or even strain rate components (as long as the strain rate values have previously been computed on the nodes):

$$\vec{v}_q = \vec{v}(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q)\vec{v}_i$$

$$T_q = T(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q)T_i$$

$$\dot{\varepsilon}_{xy,q} = \dot{\varepsilon}_{xy}(r_q, s_q) = \sum_{i=1}^{m} \mathcal{N}_i(r_q, s_q)\dot{\varepsilon}_{xy,i}$$

Chapter 8

The building blocks of the Finite Element Method

8.1 A bit of FE terminology

We introduce here some terminology for efficient element descriptions [517]:

- For triangles/tetrahedra, the designation $P_m \times P_n$ means that each component of the velocity is approximated by continuous piecewise complete Polynomials of degree m and pressure by continuous piecewise complete Polynomials of degree n. For example $P_2 \times P_1$ means

$$u^h(x,y) \sim a_1 + a_2 x + a_3 y + a_4 xy + a_5 x^2 + a_6 y^2$$

with similar approximations for v, and

$$p^h(x,y) \sim b_1 + b_2 x + b_3 y$$

Both velocity and pressure are continuous across element boundaries, and each triangular element contains 6 velocity nodes and three pressure nodes.

- For the same families, $P_m \times P_{-n}$ is as above, except that pressure is approximated via piecewise *discontinuous* polynomials of degree n. For instance, $P_2 \times P_{-1}$ is the same as $P_2 P_1$ except that pressure is now an independent linear function in each element and therefore discontinuous at element boundaries.

- For quadrilaterals/hexahedra, the designation $Q_m \times Q_n$ means that each component of the velocity is approximated by a continuous piecewise polynomial of degree m *in each direction* on the quadrilateral and likewise for pressure, except that the polynomial is of degree n. For instance, $Q_2 \times Q_1$ means

$$u^h(x,y) \sim a_1 + a_2 x + a_3 y + a_4 xy + a_5 x^2 + a_6 y^2 + a_7 x^2 y + a_8 xy^2 + a_9 x^2 y^2$$

and

$$p^h(x,y) \sim b_1 + b_2 x + b_3 y + b_4 xy$$

- For these same families, $Q_m \times Q_{-n}$ is as above, except that the pressure approximation is not continuous at element boundaries.

- Again for the same families, $Q_m \times P_{-n}$ indicates the same velocity approximation with a pressure approximation that is a discontinuous complete piecewise polynomial of degree n (not of degree n in each direction !)

- The designation P_m^+ or Q_m^+ means that some sort of bubble function was added to the polynomial approximation for the velocity. You may also find the term 'enriched element' in the literature.

- Finally, for $n = 0$, we have piecewise-constant pressure, and we omit the minus sign for simplicity.

Another point which needs to be clarified is the use of so-called 'conforming elements' (or 'non-conforming elements'). Following again Gresho & Sani [517], conforming velocity elements are those for which the basis functions form a subset of H^1 for the continuous problem (the first derivatives and their squares are integrable in Ω). For instance, the rotated $Q_1 \times P_0$ element of Rannacher and Turek (see section 10.4.9) is such that the velocity is discontinous across element edges, so that the derivative does not exist there. Another typical example of non-conforming element is the Crouzeix-Raviart element [309].

8.2 Elements and basis functions in 1D

8.2.1 Linear basis functions (Q_1)

Let $f(r)$ be a C^1 function on the interval $[-1:1]$ with $f(-1) = f_1$ and $f(1) = f_2$.
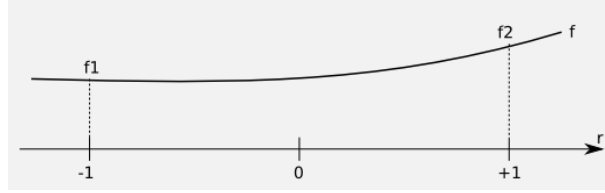
Let us assume that the function $f(r)$ is to be approximated on $[-1, 1]$ by the first order polynomial

$$f^h(r) = a + br \tag{8.1}$$

Then it must fulfil

$$
\begin{aligned}
f^h(r = -1) &= a - b = f_1 \\
f^h(r = +1) &= a + b = f_2
\end{aligned}
$$

This leads to

$$
\begin{aligned}
a &= \frac{1}{2}(f_1 + f_2) \\
b &= \frac{1}{2}(-f_1 + f_2)
\end{aligned}
\tag{8.2}
$$

and then replacing a, b in Eq. (8.1) by the above values one gets

$$f^h(r) = \left[\frac{1}{2}(1 - r)\right] f_1 + \left[\frac{1}{2}(1 + r)\right] f_2$$

or

$$f^h(r) = \sum_{i=1}^{2} N_i(r) f_1$$

with

$$
\begin{aligned}
N_1(r) &= \frac{1}{2}(1 - r) \\
N_2(r) &= \frac{1}{2}(1 + r)
\end{aligned}
\tag{8.3}
$$
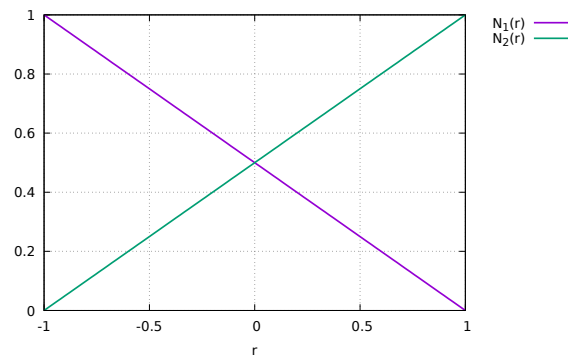
Plot of the two linear functions $N_1(r)$ and $N_2(r)$.

8.2.2   Quadratic basis functions (Q_2)

Let $f(r)$ be a C^1 function on the interval $[-1 : 1]$ with $f(-1) = f_1$, $f(0) = f_2$ and $f(1) = f_3$.
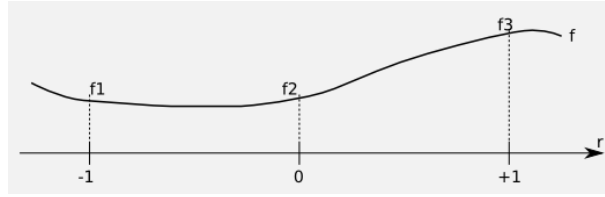
Let us assume that the function $f(r)$ is to be approximated on $[-1, 1]$ by the second order polynomial $f^h(r)$:

$$f(r) = a + br + cr^2 \tag{8.4}$$

Then it must fulfil

$$\begin{aligned} f^h(r = -1) &= a - b + c = f_1 \\ f^h(r = 0) &= a \qquad\quad = f_2 \\ f^h(r = +1) &= a + b + c = f_3 \end{aligned}$$

This leads to

$$\begin{aligned} a &= f_2 \\ b &= \frac{1}{2}(-f_1 + f_3) \\ c &= \frac{1}{2}(f_1 + f_3 - 2f_2) \end{aligned} \tag{8.5}$$

and then replacing a, b, c in Eq. (8.4) by the above values on gets

$$f^h(r) = \left[\frac{1}{2}r(r-1)\right] f_1 + (1 - r^2)f_2 + \left[\frac{1}{2}r(r+1)\right] f_3$$

or,

$$\boxed{f^h(r) = \sum_{i=1}^{3} N_i(r)f_i}$$

with

$$\begin{aligned} N_1(r) &= \frac{1}{2}r(r-1) \\ N_2(r) &= (1 - r^2) \\ N_3(r) &= \frac{1}{2}r(r+1) \end{aligned} \tag{8.6}$$
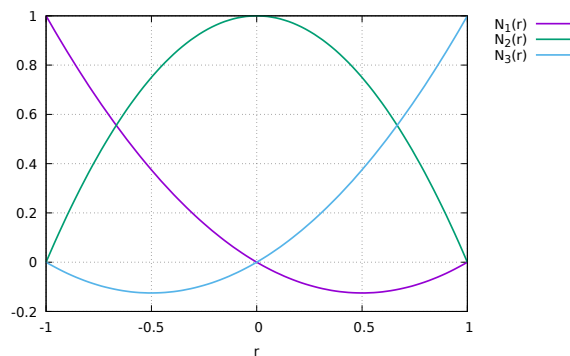
Plot of the three quadratic functions $N_1(r)$, $N_2(r)$ and $N_3(r)$.

265

Note that Q_2 basis functions can take negative values.

We will later need the first-order derivatives of these functions:

$$
\begin{aligned}
\frac{\partial N_1}{\partial r} &= r - \frac{1}{2} \\
\frac{\partial N_2}{\partial r} &= -2r \\
\frac{\partial N_3}{\partial r} &= r + \frac{1}{2}
\end{aligned}
\tag{8.7}
$$

8.2.3 Cubic basis functions (Q_3)

We proceed as previously by assuming that the third-order polynomial representation of function $f(r)$ is given by

$$
f^h(r) = a + br + cr^2 + dr^3
$$

with the nodes at position -1,-1/3, +1/3 and +1. It then must fulfil all four conditions:

$$
\begin{aligned}
f(-1) &= a - b + c - d = f_1 \\
f(-1/3) &= a - \frac{b}{3} + \frac{c}{9} - \frac{d}{27} = f_2 \\
f(+1/3) &= a - \frac{b}{3} + \frac{c}{9} - \frac{d}{27} = f_3 \\
f(+1) &= a + b + c + d = f_4
\end{aligned}
$$

Adding the first and fourth equation and the second and third, one arrives at

$$
f_1 + f_4 = 2a + 2c \qquad f_2 + f_3 = 2a + \frac{2c}{9}
$$

and finally:

$$
a = \frac{1}{16}\left(-f_1 + 9f_2 + 9f_3 - f_4\right)
$$

$$
c = \frac{9}{16}\left(f_1 - f_2 - f_3 + f_4\right)
$$

Combining the original 4 equations in a different way yields

$$
2b + 2d = f_4 - f_1 \qquad \frac{2b}{3} + \frac{2d}{27} = f_3 - f_2
$$

so that

$$
b = \frac{1}{16}\left(f_1 - 27f_2 + 27f_3 - f_4\right)
$$

$$
d = \frac{9}{16}\left(-f_1 + 3f_2 - 3f_3 + f_4\right)
$$

Finally,

$$
\begin{aligned}
f^h(r) &= a + b + cr^2 + dr^3 \\
&= \frac{1}{16}(-1 + r + 9r^2 - 9r^3)f_1 \\
&+ \frac{1}{16}(9 - 27r - 9r^2 + 27r^3)f_2 \\
&+ \frac{1}{16}(9 + 27r - 9r^2 - 27r^3)f_3 \\
&+ \frac{1}{16}(-1 - r + 9r^2 + 9r^3)f_4 \\
&= \sum_{i=1}^{4} N_i(r)f_i
\end{aligned}
$$

where (see also for example [825, p49])

$$
\begin{aligned}
N_1 &= \frac{1}{16}(-1 + r + 9r^2 - 9r^3) \\
N_2 &= \frac{1}{16}(9 - 27r - 9r^2 + 27r^3) \\
N_3 &= \frac{1}{16}(9 + 27r - 9r^2 - 27r^3) \\
N_4 &= \frac{1}{16}(-1 - r + 9r^2 + 9r^3)
\end{aligned}
$$
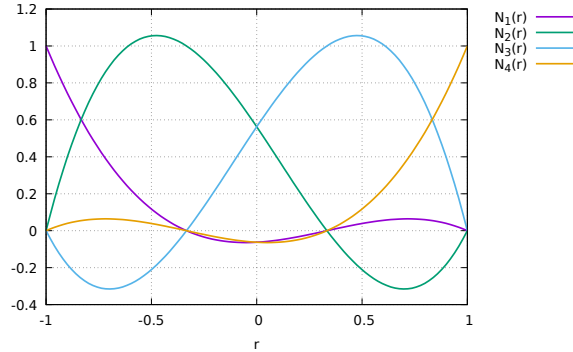
Plot of the four cubic functions $N_1(r)$, $N_2(r)$, $N_3(r)$ and $N_4(r)$.

Let us now verify that these functions can represent any polynomial function up to third order:

- Let us assume $f(r) = C$, then

$$
f^h(r) = \sum N_i(r)f_i = \sum_i N_i C = C \sum_i N_i = C
$$

so that a constant function is exactly reproduced, as expected. This is a very important property of the N_i functions: They must fulfil $\sum_i N_i = 1$.

- Let us assume $f(r) = r$, then $f_1 = -1$, $f_2 = -1/3$, $f_3 = 1/3$ and $f_4 = +1$. We then have

$$
\begin{aligned}
f^h(r) &= \sum N_i(r) f_i \\
&= -N_1(r) - \frac{1}{3} N_2(r) + \frac{1}{3} N_3(r) + N_4(r) \\
&= [-(-1 + r + 9r^2 - 9r^3) \\
&\quad -\frac{1}{3}(9 - 27r - 9r^2 - 27r^3) \\
&\quad +\frac{1}{3}(9 + 27r - 9r^2 + 27r^3) \\
&\quad +(-1 - r + 9r^2 + 9r^3)]/16 \\
&= [-r + 9r + 9r - r]/16 + ...0... \\
&= r
\end{aligned}
\tag{8.8}
$$

- The cases $f(r) = r^2$ and $f(r) = r^3$ are left as exercise.

The basis functions first-order derivatives are given by

$$
\begin{aligned}
\frac{\partial N_1}{\partial r} &= \frac{1}{16}(1 + 18r - 27r^2) \\
\frac{\partial N_2}{\partial r} &= \frac{1}{16}(-27 - 18r + 81r^2) \\
\frac{\partial N_3}{\partial r} &= \frac{1}{16}(+27 - 18r - 81r^2) \\
\frac{\partial N_4}{\partial r} &= \frac{1}{16}(-1 + 18r + 27r^2)
\end{aligned}
$$

We can also verify that the derivatives are also properly approximated:

- Let us assume $f(r) = C$, then

$$
\begin{aligned}
\frac{\partial f^h}{\partial r} &= \sum_i \frac{\partial N_i}{\partial r} f_i \\
&= C \sum_i \frac{\partial N_i}{\partial r} \\
&= \frac{C}{16}[(1 + 18r - 27r^2) + (-27 - 18r + 81r^2) + (+27 - 18r - 81r^2) + (-1 + 18r + 27r^2)] \\
&= 0
\end{aligned}
$$

- Let us assume $f(r) = r$, then $f_1 = -1$, $f_2 = -1/3$, $f_3 = 1/3$ and $f_4 = +1$. We then have

$$
\begin{aligned}
\frac{\partial f^h}{\partial r} &= \sum_i \frac{\partial N_i}{\partial r} f_i \\
&= \frac{1}{16}[-(1 + 18r - 27r^2) - \frac{1}{3}(-27 - 18r + 81r^2) + \frac{1}{3}(27 - 18r - 81r^2) + (-1 + 18r + 27r^2)] \\
&= \frac{1}{16}[-2 + 18 + 54r^2 - 54r^2] \\
&= 1
\end{aligned}
$$

- Let us assume $f(r) = r^2$, then $f_1 = 1$, $f_2 = 1/9$, $f_3 = 1/9$ and $f_4 = 1$. We then have

$$
\begin{aligned}
\frac{\partial f^h}{\partial r} &= \sum_i \frac{\partial N_i}{\partial r} f_i \\
&= \frac{1}{16}\left[(1 + 18r - 27r^2) + \frac{1}{9}(-27 - 18r + 81r^2) + \frac{1}{9}(27 - 18r - 81r^2) + (-1 + 18r + 27r^2)\right] \\
&= \frac{1}{16}(32r) \\
&= 2r
\end{aligned}
\tag{8.9}
$$

as expected.

8.2.4 Quartic basis functions (Q_4)

The 1D basis polynomial is given by

$$f_h(r) = a + br + cr^2 + dr^3 + er^4$$

with the nodes at position -1,-1/2, 0, +1/2 and +1. The function $f^h(r)$ must then fulfil

$$
\begin{aligned}
f_h(-1) &= a - b + c - d + e = f_1 \\
f_h(-1/2) &= a - \frac{b}{2} + \frac{c}{4} - \frac{d}{8} + \frac{e}{16} = f_2 \\
f_h(0) &= a = f_3 \\
f_h(+1/2) &= a - \frac{b}{2} + \frac{c}{4} - \frac{d}{8} + \frac{e}{16} = f_4 \\
f_h(+1) &= a + b + c + d + e = f_5
\end{aligned}
$$

or,

$$
\begin{pmatrix}
1 & -1 & 1 & -1 & 1 \\
1 & -1/2 & 1/4 & -1/8 & 1/16 \\
1 & 0 & 0 & 0 & 0 \\
1 & 1/2 & 1/4 & 1/8 & 1/16 \\
1 & 1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix}
=
\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix}
\tag{8.10}
$$

The third line gives $a = f_3$ so that

$$
\underbrace{
\begin{pmatrix}
-1 & 1 & -1 & 1 \\
-1/2 & 1/4 & -1/8 & 1/16 \\
1/2 & 1/4 & 1/8 & 1/16 \\
1 & 1 & 1 & 1
\end{pmatrix}
}_{A}
\begin{pmatrix} b \\ c \\ d \\ e \end{pmatrix}
=
\begin{pmatrix} f_1 - f_3 \\ f_2 - f_3 \\ f_4 - f_3 \\ f_5 - f_3 \end{pmatrix}
\tag{8.11}
$$

The inverse of the matrix A is:

$$
A^{-1} = \frac{1}{6}
\begin{pmatrix}
1 & -8 & 8 & -1 \\
-1 & 16 & 16 & -1 \\
-4 & 8 & -8 & 4 \\
4 & -16 & -16 & 4
\end{pmatrix}
$$

so that

$$
\begin{pmatrix} b \\ c \\ d \\ e \end{pmatrix}
= \frac{1}{6}
\begin{pmatrix}
1 & -8 & 8 & -1 \\
-1 & 16 & 16 & -1 \\
-4 & 8 & -8 & 4 \\
4 & -16 & -16 & 4
\end{pmatrix}
\cdot
\begin{pmatrix} f_1 - f_3 \\ f_2 - f_3 \\ f_4 - f_3 \\ f_5 - f_3 \end{pmatrix}
$$

and then

$$b = \frac{1}{6}(f_1 - 8f_2 + 8f_4 - f_5) \tag{8.12}$$

$$c = \frac{1}{6}(-f_1 + 16f_2 - 30f_3 + 16f_4 - f_5) \tag{8.13}$$

$$d = \frac{1}{6}(-4f_1 + 8f_2 - 8f_4 + 4f_5) \tag{8.14}$$

$$e = \frac{1}{6}(4f_1 - 16f_2 + 24f_3 - 16f_4 + 4f_5) \tag{8.15}$$

Finally

$$
\begin{aligned}
f_h(r) &= a + br + cr^2 + dr^3 + er^4 \\
&= f_3 + \frac{1}{6}(f_1 - 8f_2 + 8f_4 - f_5)r + \frac{1}{6}(-f_1 + 16f_2 - 30f_3 + 16f_4 - f_5)r^2 + \\
&\quad \frac{1}{6}(-4f_1 + 8f_2 - 8f_4 + 4f_5)r^3 + \frac{1}{6}(4f_1 - 16f_2 + 24f_3 - 16f_4 + 4f_5)r^4 \\
&= \frac{1}{6}(r - r^2 - 4r^3 + 4r^4)f_1 \\
&+ \frac{1}{6}(-8r + 16r^2 + 8r^3 - 16r^4)f_2 \\
&+ (1 - 5r^2 + 4r^4)f_3 \\
&+ \frac{1}{6}(8r + 16r^2 - 8r^3 - 16r^4)f_4 \\
&+ \frac{1}{6}(-r - r^2 + 4r^3 + 4r^4)f_5
\end{aligned}
$$

with

$$
\begin{aligned}
N_1(r) &= \frac{1}{6}(r - r^2 - 4r^3 + 4r^4) \\
N_2(r) &= \frac{1}{6}(-8r + 16r^2 + 8r^3 - 16r^4) \\
N_3(r) &= (1 - 5r^2 + 4r^4) \\
N_4(r) &= \frac{1}{6}(8r + 16r^2 - 8r^3 - 16r^4) \\
N_5(r) &= \frac{1}{6}(-r - r^2 + 4r^3 + 4r^4)
\end{aligned}
\tag{8.16}
$$
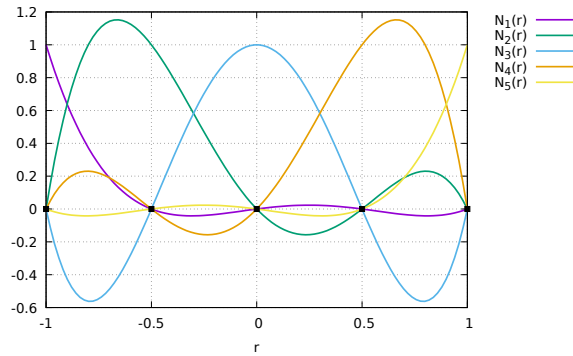
Plot of the 5 quartic basis functions.

The basis functions derivative are given by

$$\frac{\partial N_1}{\partial r} = \frac{1}{6}(1 - 2r - 12r^2 + 16r^3)$$

$$\frac{\partial N_2}{\partial r} = \frac{1}{6}(-8 + 32r + 24r^2 - 64r^3)$$

$$\frac{\partial N_3}{\partial r} = -10r + 16r^3$$

$$\frac{\partial N_4}{\partial r} = \frac{1}{6}(8 + 32r - 24r^2 - 64r^3)$$

$$\frac{\partial N_5}{\partial r} = \frac{1}{6}(-1 - 2r + 12r^2 + 16r^3) \tag{8.17}$$

8.2.5 Fifth-order basis functions (Q_5)

Following the methodology presented hereafter for Q_6, we arrive at

$$\mathcal{N}_1(r) = -\frac{625}{768}(r + \frac{3}{5})(r + \frac{1}{5})(r - \frac{1}{5})(r - \frac{3}{5})(r - 1)$$

$$\mathcal{N}_2(r) = \frac{3125}{768}(r + 1)(r + \frac{1}{5})(r - \frac{1}{5})(r - \frac{3}{5})(r - 1)$$

$$\mathcal{N}_3(r) = -\frac{3125}{384}(r + 1)(r + \frac{3}{5})(r - \frac{1}{5})(r - \frac{3}{5})(r - 1)$$

$$\mathcal{N}_4(r) = \frac{3125}{384}(r + 1)(r + \frac{3}{5})(r + \frac{1}{5})(r - \frac{3}{5})(r - 1)$$

$$\mathcal{N}_5(r) = -\frac{3125}{768}(r + 1)(r + \frac{3}{5})(r + \frac{1}{5})(r - \frac{1}{5})(r - 1)$$

$$\mathcal{N}_6(r) = \frac{625}{768}(r + 1)(r + \frac{3}{5})(r + \frac{1}{5})(r - \frac{1}{5})(r - \frac{3}{5}) \tag{8.18}$$

or,

$$\mathcal{N}_1(r) = -\frac{1}{768}(625r^5 - 625r^4 - 250r^3 + 250r^2 + 9r - 9)$$

$$\mathcal{N}_2(r) = \frac{25}{768}(125r^5 - 75r^4 - 130r^3 + 78r^2 + 5r - 3)$$

$$\mathcal{N}_3(r) = -\frac{25}{384}(125r^5 - 25r^4 - 170r^3 + 34r^2 + 45r - 9)$$

$$\mathcal{N}_4(r) = \frac{25}{384}(125r^5 + 25r^4 - 170r^3 - 34r^2 + 45r + 9)$$

$$\mathcal{N}_5(r) = -\frac{25}{768}(125r^5 + 75r^4 - 130r^3 - 78r^2 + 5r + 3)$$

$$\mathcal{N}_6(r) = \frac{1}{768}(625r^5 + 625r^4 - 250r^3 - 250r^2 + 9r + 9) \tag{8.19}$$

with the derivatives given by

$$\frac{\partial N_1}{\partial r} = -\frac{1}{768}(3125r^4 - 2500r^3 - 750r^2 + 500r + 9)$$

$$\frac{\partial N_2}{\partial r} = \frac{25}{768}(625r^4 - 300r^3 - 390r^2 + 156r + 5)$$

$$\frac{\partial N_3}{\partial r} = -\frac{25}{384}(625r^4 - 100r^3 - 510r^2 + 68r + 45)$$

$$\frac{\partial N_4}{\partial r} = \frac{25}{384}(625r^4 + 100r^3 - 510r^2 - 68r + 45)$$

$$\frac{\partial N_5}{\partial r} = -\frac{25}{768}(625r^4 + 300r^3 - 390r^2 - 156r + 5)$$

$$\frac{\partial N_6}{\partial r} = \frac{1}{768}(3125r^4 + 2500r^3 - 750r^2 - 500r + 9) \tag{8.20}$$
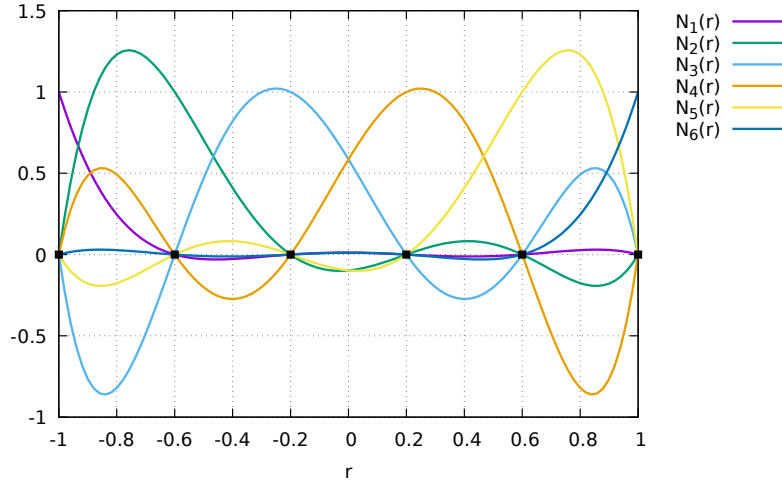
Plot of the 6 fifth-order basis functions.

These functions are used in STONE **??**.

8.2.6 Sixth-order basis functions (Q_6)

The 1D basis polynomial is given by

$$f_h(r) = a + br + cr^2 + dr^3 + er^4 + fr^5 + gr^6$$

with the nodes at position -1,-2/3, -1/3, 0, +1/3, +2/3 and +1. The function $f^h(r)$ must then fulfil

$$\begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -\frac{2}{3} & \frac{4}{9} & -\frac{8}{27} & \frac{16}{81} & -\frac{32}{243} & \frac{64}{729} \\ 1 & -\frac{1}{3} & \frac{1}{9} & -\frac{1}{27} & \frac{1}{81} & -\frac{1}{243} & \frac{1}{729} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & \frac{1}{3} & \frac{1}{9} & \frac{1}{27} & \frac{1}{81} & \frac{1}{243} & \frac{1}{729} \\ 1 & \frac{2}{3} & \frac{4}{9} & \frac{8}{27} & \frac{16}{81} & \frac{32}{243} & \frac{64}{729} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \end{pmatrix}$$

The middle line yields $a = f_4$, so that we have:

$$
\begin{pmatrix}
-1 & 1 & -1 & 1 & -1 & 1 \\
-\frac{2}{3} & \frac{4}{9} & -\frac{8}{27} & \frac{16}{81} & -\frac{32}{243} & \frac{64}{729} \\
-\frac{1}{3} & \frac{1}{9} & -\frac{1}{27} & \frac{1}{81} & -\frac{1}{243} & \frac{1}{729} \\
\frac{1}{3} & \frac{1}{9} & \frac{1}{27} & \frac{1}{81} & \frac{1}{243} & \frac{1}{729} \\
\frac{2}{3} & \frac{4}{9} & \frac{8}{27} & \frac{16}{81} & \frac{32}{243} & \frac{64}{729} \\
1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\cdot
\begin{pmatrix} b \\ c \\ d \\ e \\ f \\ g \end{pmatrix}
=
\begin{pmatrix} f_1 - f_4 \\ f_2 - f_4 \\ f_3 - f_4 \\ f_5 - f_4 \\ f_6 - f_4 \\ f_7 - f_4 \end{pmatrix}
$$

Multiplying all lines by 729, we obtain:

$$
\frac{1}{729}
\begin{pmatrix}
-729 & 729 & -729 & 729 & -729 & 729 \\
-486 & 324 & -216 & 144 & -96 & 64 \\
-243 & 81 & -27 & 9 & -3 & 1 \\
243 & 81 & 27 & 9 & 3 & 1 \\
486 & 324 & 216 & 144 & 96 & 64 \\
729 & 729 & 729 & 729 & 729 & 729
\end{pmatrix}
\cdot
\begin{pmatrix} b \\ c \\ d \\ e \\ f \\ g \end{pmatrix}
=
\begin{pmatrix} f_1 - f_4 \\ f_2 - f_4 \\ f_3 - f_4 \\ f_5 - f_4 \\ f_6 - f_4 \\ f_7 - f_4 \end{pmatrix}
$$

The inverse[1] of this matrix is:

```
-0.00006859   0.00061728  -0.00308642   0.00308642  -0.00061728   0.00006859
0.00006859   -0.00092593   0.00925926   0.00925926  -0.00092593   0.00006859
0.00077160   -0.00617284   0.01003086  -0.01003086   0.00617284  -0.00077160
-0.00077160   0.00925926  -0.03009259  -0.03009259   0.00925926  -0.00077160
-0.00138889   0.00555556  -0.00694444   0.00694444  -0.00555556   0.00138889
0.00138889   -0.00833333   0.02083333   0.02083333  -0.00833333   0.00138889
```

Obviously, this is not a very practical approach anymore. One could solve the system by hand, making sure to keep fractions but it will be cumbersome. Let us turn to another approach.

The nodes inside the reference element are as follows:

```
(1) (2)   (3) (4) (5)   (6) (7)
-|---|----|---+---|----|---|-
-1 -2/3 -1/3  0  1/3  2/3  1
```

Basis function $\mathcal{N}_1(r)$ is a 6th order polynomial expression that should be 1 at node 1, and 0 at others, i.e. at $r = -2/3, -1/3, 0, 1/3, 2/3, 1$. It must then be of the form:

$$
\mathcal{N}_1(r) = \alpha(r + \frac{2}{3})(r + \frac{1}{3})(r)(r - \frac{1}{3})(r - \frac{2}{3})(r - 1)
$$

When evaluated at $r = -1$, we get

$$
\mathcal{N}_1(r = -1) = \alpha(-\frac{1}{3})(-\frac{2}{3})(-1)(-\frac{4}{3})(-\frac{5}{3})(-2) = \alpha\frac{80}{81}
$$

Since this quantity must be 1, we have

$$
1 = \alpha\frac{80}{81} \quad \rightarrow \quad \alpha = \frac{81}{80}
$$

so that

$$
\begin{aligned}
\mathcal{N}_1(r) &= \frac{81}{80}(r + \frac{2}{3})(r + \frac{1}{3})(r)(r - \frac{1}{3})(r - \frac{2}{3})(r - 1) \\
&= \frac{81}{80}\frac{1}{81}(3r + 2)(3r + 1)(r)(3r - 1)(3r - 2)(r - 1) \\
&= \frac{1}{80}(9r^2 - 4)(9r^2 - 1)(r^2 - r) \\
&= \frac{1}{80}(81r^4 - 45r^2 + 4)(r^2 - r) \quad\quad (8.21)
\end{aligned}
$$

---

[1]https://physandmathsolutions.com/Matrices/matrix_inverse/matrix_inverse_6x6.php

Moving to $\mathcal{N}_2(r)$, we have

$$\mathcal{N}_2(r) = \alpha(r+1)(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1)$$

which must be equal to 1 for $r = -2/3$:

$$
\begin{aligned}
\mathcal{N}_2(r=-2/3) &= \alpha(-\frac{2}{3}+1)(-\frac{2}{3}+\frac{1}{3})(-\frac{2}{3})(-\frac{2}{3}-\frac{1}{3})(-\frac{2}{3}-\frac{2}{3})(-\frac{2}{3}-1) \\
&= \alpha(\frac{1}{3})(-\frac{1}{3})(-\frac{2}{3})(-1)(-\frac{4}{3})(-\frac{5}{3}) \\
&= -\alpha\frac{40}{243}
\end{aligned}
\tag{8.22}
$$

so that

$$\mathcal{N}_2(r) = -\frac{243}{40}(r+1)(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1)$$

Moving to $\mathcal{N}_3(r)$, we have

$$\mathcal{N}_3(r) = \alpha(r+1)(r+\frac{2}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1)$$

which must be equal to 1 for $r = -1/3$:

$$
\begin{aligned}
\mathcal{N}_3(r=-1/3) &= \alpha(-\frac{1}{3}+1)(-\frac{1}{3}+\frac{2}{3})(-\frac{1}{3})(-\frac{1}{3}-\frac{1}{3})(-\frac{1}{3}-\frac{2}{3})(-\frac{1}{3}-1) \\
&= \alpha(\frac{2}{3})(\frac{1}{3})(-\frac{1}{3})(-\frac{2}{3})(-1)(-\frac{4}{3}) \\
&= \alpha\frac{16}{243}
\end{aligned}
\tag{8.23}
$$

so that

$$\mathcal{N}_3(r) = \frac{243}{16}(r+1)(r+\frac{2}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1)$$

Likewise, we arrive at the rest of the basis functions. In the end:

$$
\begin{aligned}
\mathcal{N}_1(r) &= \frac{81}{80}(r+\frac{2}{3})(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1) \\
\mathcal{N}_2(r) &= -\frac{243}{40}(r+1)(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1) \\
\mathcal{N}_3(r) &= \frac{243}{16}(r+1)(r+\frac{2}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})(r-1) \\
\mathcal{N}_4(r) &= -\frac{81}{4}(r+1)(r+\frac{2}{3})(r+\frac{1}{3})(r-\frac{1}{3})(r-\frac{2}{3})(r-1) \\
\mathcal{N}_5(r) &= \frac{243}{16}(r+1)(r+\frac{2}{3})(r+\frac{1}{3})(r)(r-\frac{2}{3})(r-1) \\
\mathcal{N}_6(r) &= -\frac{243}{40}(r+1)(r+\frac{2}{3})(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-1) \\
\mathcal{N}_7(r) &= \frac{81}{80}(r+1)(r+\frac{2}{3})(r+\frac{1}{3})(r)(r-\frac{1}{3})(r-\frac{2}{3})
\end{aligned}
\tag{8.24}
$$

or

$$\mathcal{N}_1(r) = \frac{1}{80}(81r^6 - 81r^5 - 45r^4 + 45r^3 + 4r^2 - 4r)$$

$$\mathcal{N}_2(r) = -\frac{9}{40}(27r^6 - 18r^5 - 30r^4 + 20r^3 + 3r^2 - 2r)$$

$$\mathcal{N}_3(r) = \frac{9}{16}(27r^6 - 9r^5 - 39r^4 + 13r^3 + 12r^2 - 4r)$$

$$\mathcal{N}_4(r) = -\frac{1}{4}(81r^6 - 126r^4 + 49r^2 - 4)$$

$$\mathcal{N}_5(r) = \frac{9}{16}(27r^6 + 9r^5 - 39r^4 - 13r^3 + 12r^2 + 4r)$$

$$\mathcal{N}_6(r) = -\frac{9}{40}(27r^6 + 18r^5 - 30r^4 - 20r^3 + 3r^2 + 2r)$$

$$\mathcal{N}_7(r) = \frac{1}{80}(81r^6 + 81r^5 - 45r^4 - 45r^3 + 4r^2 + 4r) \tag{8.25}$$
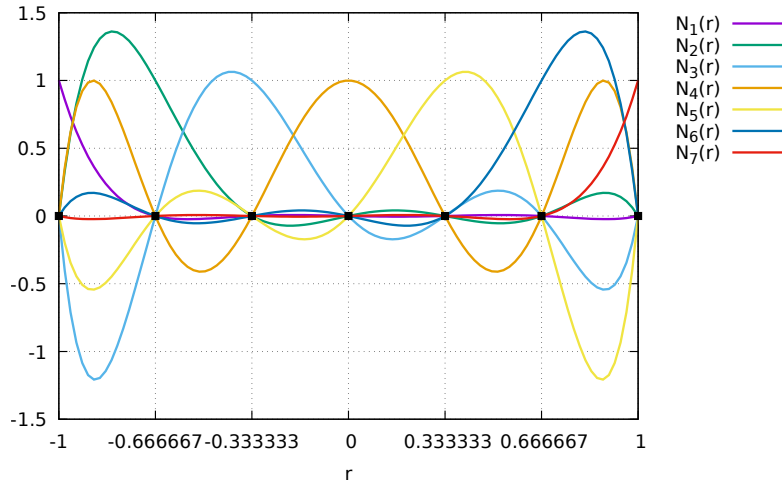
Plot of the 7 six-order basis functions.

Using WolframAlpha[2], we arrive at

$$\frac{d\mathcal{N}_1}{dr} = \frac{1}{80}(486r^5 - 405r^4 - 180r^3 + 135r^2 + 8r - 4)$$

$$\frac{d\mathcal{N}_2}{dr} = -\frac{9}{20}(81r^5 - 45r^4 - 60r^3 + 30r^2 + 3r - 1)$$

$$\frac{d\mathcal{N}_3}{dr} = \frac{9}{16}(162r^5 - 45r^4 - 156r^3 + 39r^2 + 24r - 4)$$

$$\frac{d\mathcal{N}_4}{dr} = \frac{1}{2}(-243r^5 + 252r^3 - 49r)$$

$$\frac{d\mathcal{N}_5}{dr} = \frac{9}{16}(162r^5 + 45r^4 - 156r^3 - 39r^2 + 24r + 4)$$

$$\frac{d\mathcal{N}_6}{dr} = -\frac{9}{20}(81r^5 + 45r^4 - 60r^3 - 30r^2 + 3r + 1)$$

$$\frac{d\mathcal{N}_7}{dr} = \frac{1}{80}(486r^5 + 405r^4 - 180r^3 - 135r^2 + 8r + 4) \tag{8.26}$$

These functions are used in STONE f152.

8.2.7 A generic approach to 1D basis functions

In order to define basis functions of order n each element must have $n + 1$ nodes. The $i - th$ basis function for an $n - th$ order approximation is given by:

$$\mathcal{N}_i(r) = \frac{\prod_{j=0, j \neq i}^{n}(r - r_j)}{\prod_{j=0, j \neq i}^{n}(r_i - r_j)}$$

Let us see in practice how this works and start with $n = 2$ (i.e. Q_2 basis functions). In the reference element we have $r_0 = -1$, $r_1 = 0$ and $r_2 = +1$, so that

$$
\begin{aligned}
\mathcal{N}_0 &= \frac{(r - r_1)(r - r_2)}{(r_0 - r_1)(r_0 - r_2)} \\
&= \frac{(r - 0)(r - 1)}{(-1 - 0)(-1 - 1)} \\
&= \frac{1}{2}r(r - 1) \\
\mathcal{N}_1 &= \frac{(r - r_0)(r - r_2)}{(r_1 - r_0)(r_1 - r_2)} \\
&= \frac{(r + 1)(r - 1)}{(0 + 1)(0 - 1)} \\
&= 1 - r^2 \\
\mathcal{N}_2 &= \frac{(r - r_0)(r - r_1)}{(r_2 - r_0)(r_2 - r_1)} \\
&= \frac{(r + 1)(r - 0)}{(1 + 1)(1 - 0)} \\
&= \frac{1}{2}r(r + 1)
\end{aligned}
\tag{8.27}
$$

These are the basis functions obtained in Section 8.2.2.

Note that in practice it is often rarely desirable to use much higher than quadratic basis functions because higher order ones have too much oscillation.

According to Guermond[3]: "While the choice of equidistant nodes appears somewhat natural, it is appro- priate only when working with low-degree polynomials. The main difficulty comes from the oscillatory nature of the Lagrange polynomials as the num- ber of interpolation nodes grows. This phenomenon is often referred to as the Runge phenomenon [359] (see also Meray [313]".

---

[3] https://people.tamu.edu/~guermond/M661_FALL_2015/chap2.pdf

8.3 Elements and basis functions in 2D

Let us for a moment consider a single quadrilateral element in the xy-plane, as shown on the following figure:
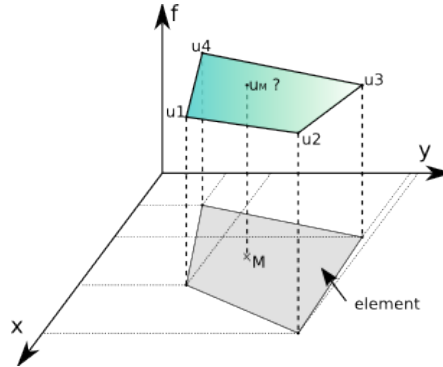
Let us assume that we know the values of a given field u at the four vertices. For a given point M inside the element in the plane, what is the value of the field u at this point? It makes sense to postulate that u_M will be given by

$$u_M = \phi(u_1, u_2, u_3, u_4, x_M, y_M)$$

where ϕ is a function to be determined. Although ϕ is not unique, we can decide to express the value u_M as a weighed sum of the values at the vertices u_i. One option could be to assign all four vertices the same weight, say $1/4$ so that $u_M = (u_1 + u_2 + u_3 + u_4)/4$, i.e. u_M is simply given by the arithmetic mean of the vertices values. If the function $u(x, y)$ is such that it is a constant function, say $u(x, y) = C$, then $u_M = (u_1 + u_2 + u_3 + u_4)/4 = (C + C + C + C)/4 = C$ and the result is exact. However, for any other function u the value u_M will not be as accurate. Also, this approach suffers from a major drawback as it does not use the location of point M inside the element. For instance, when $(x_M, y_M) \to (x_2, y_2)$ we expect $u_M \to u_2$ but u_M would remain equal to $(u_1 + u_2 + u_3 + u_4)/4$.

In light of this, we could now assume that the weights would depend on the position of M in a continuous fashion:

$$u(x_M, y_M) = \sum_{i=1}^{4} \mathcal{N}_i(x_M, y_M)\, u_i = \mathcal{N}_1(x_M, y_M)u_1 + \mathcal{N}_2(x_M, y_M)u_2 + \mathcal{N}_3(x_M, y_M)u_3 + \mathcal{N}_4(x_M, y_M)u_4$$

(8.28)

where the bN_i are continuous (and also "well behaved") functions which have the property:

$$\mathcal{N}_i(x_j, y_j) = \delta_{ij}$$

or, in other words for example:

$$
\begin{aligned}
\mathcal{N}_3(x_1, y_1) &= 0 \\
\mathcal{N}_3(x_2, y_2) &= 0 \\
\mathcal{N}_3(x_3, y_3) &= 1 \\
\mathcal{N}_3(x_4, y_4) &= 0
\end{aligned}
$$

(8.29)

The functions $\mathcal{N}_i$ are commonly called basis functions.

Omitting the M subscripts (yet stil assuming the point being inside the element), the velocity components u and v for a point inside the element are given by:

$$u^h(x,y) \;=\; \sum_{i=1}^{4} \mathcal{N}_i(x,y)\, u_i \tag{8.30}$$

$$v^h(x,y) \;=\; \sum_{i=1}^{4} \mathcal{N}_i(x,y)\, v_i \tag{8.31}$$

where we have added the superscript h to denote that it is an approximation of the functions of this element of diameter h (by diameter we mean here a representative scalar value of the dimension of the element).

One can now easily compute velocity gradients (and therefore the strain rate tensor) since we have assumed the basis functions to be "well behaved" (in this case first-order differentiable):

$$\dot{\varepsilon}^h_{xx}(x,y) \;=\; \frac{\partial u^h}{\partial x} = \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x}\, u_i \tag{8.32}$$

$$\dot{\varepsilon}^h_{yy}(x,y) \;=\; \frac{\partial v^h}{\partial y} = \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y}\, v_i \tag{8.33}$$

$$\dot{\varepsilon}^h_{xy}(x,y) \;=\; \frac{1}{2}\left(\frac{\partial u^h}{\partial y} + \frac{\partial v^h}{\partial x}\right) = \frac{1}{2}\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y}\, u_i + \frac{1}{2}\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x}\, v_i \tag{8.34}$$

How we actually obtain the exact form of the basis functions $\mathcal{N}_i$ is explained in the coming sections.

8.3.1 Bilinear basis functions (Q_1)

basis_Q1_2D.tex

In this section, we consider for simplicity an element which is a square defined by $-1 < r < 1$, $-1 < s < 1$ in the Cartesian coordinates system (r,s)[4]:

(tikz_q12d.tex)
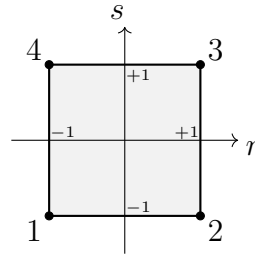
Note the counter-clockwise numbering[5]. This element is commonly called the reference element. How we go from the (x,y) coordinate system to the (r,s) once and vice versa will be dealt with later on. The basis functions in the above reference element in the reduced coordinates system (r,s) are given by:

---

[4]There is a reason to choose r and s as coordinates and not x and y as we will see later.
[5]Note that in many of the python codes which are part of this project the numbering starts at 0.

$$
\begin{aligned}
\mathcal{N}_1(r,s) &= 0.25(1-r)(1-s) \\
\mathcal{N}_2(r,s) &= 0.25(1+r)(1-s) \\
\mathcal{N}_3(r,s) &= 0.25(1+r)(1+s) \\
\mathcal{N}_4(r,s) &= 0.25(1-r)(1+s)
\end{aligned} \tag{8.35}
$$

These basis functions are the product of the linear basis functions of Section 8.2.1 in the r direction and the s direction.
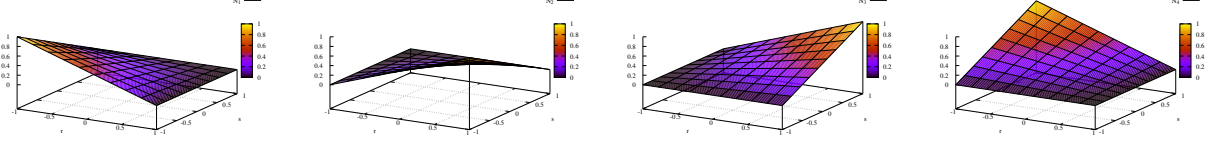
Surface representation of the basis functions on the reference element. in images/basis_Q1_2D/

The partial derivatives of these functions with respect to r ans s automatically follow:

$$
\begin{aligned}
\frac{\partial \mathcal{N}_1}{\partial r}(r,s) &= -0.25(1-s) & \qquad \frac{\partial \mathcal{N}_1}{\partial s}(r,s) &= -0.25(1-r) \\
\frac{\partial \mathcal{N}_2}{\partial r}(r,s) &= +0.25(1-s) & \frac{\partial \mathcal{N}_2}{\partial s}(r,s) &= -0.25(1+r) \\
\frac{\partial \mathcal{N}_3}{\partial r}(r,s) &= +0.25(1+s) & \frac{\partial \mathcal{N}_3}{\partial s}(r,s) &= +0.25(1+r) \\
\frac{\partial \mathcal{N}_4}{\partial r}(r,s) &= -0.25(1+s) & \frac{\partial \mathcal{N}_4}{\partial s}(r,s) &= +0.25(1-r)
\end{aligned}
$$

Let us go back to Eq. (8.31) and let us assume that the function $v(r,s) = C$ so that $v_i = C$ for $i = 1, 2, 3, 4$. It then follows that

$$
v^h(r,s) = \sum_{i=1}^{4} \mathcal{N}_i(r,s)\, v_i = C \sum_{i=1}^{4} \mathcal{N}_i(r,s) = C[\mathcal{N}_1(r,s) + \mathcal{N}_2(r,s) + \mathcal{N}_3(r,s) + \mathcal{N}_4(r,s)] = C
$$

This is a very important property: if the v function used to assign values at the vertices is constant, then the value of v^h *anywhere* in the element is exactly C. If we now turn to the derivatives of v with respect to r and s:

$$
\frac{\partial v^h}{\partial r}(r,s) = \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\, v_i = C \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) = C\left[-0.25(1-s) + 0.25(1-s) + 0.25(1+s) - 0.25(1+s)\right] =
$$

$$
\frac{\partial v^h}{\partial s}(r,s) = \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\, v_i = C \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = C\left[-0.25(1-r) - 0.25(1+r) + 0.25(1+r) + 0.25(1-r)\right] =
$$

We reassuringly find that the derivative of a constant field anywhere in the element is exactly zero.

If we now choose $v(r, s) = ar + bs$ with a and b two constant scalars, we find:

$$
\begin{aligned}
v^h(r, s) &= \sum_{i=1}^{4} \mathcal{N}_i(r, s) \, v_i \\
&= \sum_{i=1}^{4} \mathcal{N}_i(r, s)(ar_i + bs_i) \\
&= a \sum_{i=1}^{4} \mathcal{N}_i(r, s)r_i + b \sum_{i=1}^{4} \mathcal{N}_i(r, s)s_i \\
&= a \left[\frac{1}{4}(1 - r)(1 - s)(-1) + \frac{1}{4}(1 + r)(1 - s)(+1) + \frac{1}{4}(1 + r)(1 + s)(+1) + \frac{1}{4}(1 - r)(1 + s)(-1) \right] \\
&+ b \left[\frac{1}{4}(1 - r)(1 - s)(-1) + \frac{1}{4}(1 + r)(1 - s)(-1) + \frac{1}{4}(1 + r)(1 + s)(+1) + \frac{1}{4}(1 - r)(1 + s)(+1) \right] \\
&= \frac{a}{4} \left[-(1 - r)(1 - s) + (1 + r)(1 - s) + (1 + r)(1 + s) - (1 - r)(1 + s) \right] \\
&+ \frac{b}{4} \left[-(1 - r)(1 - s) - (1 + r)(1 - s) + (1 + r)(1 + s) + (1 - r)(1 + s) \right] \\
&= ar + bs
\end{aligned}
\tag{8.36}
$$

This set of bilinear basis functions is therefore capable of exactly representing a bilinear field. The derivatives are:

$$
\begin{aligned}
\frac{\partial v^h}{\partial r}(r, s) &= \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial r}(r, s) \, v_i \\
&= a \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial r}(r, s)r_i + b \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial r}(r, s)s_i \\
&= a \left[-\frac{1}{4}(1 - s)(-1) + \frac{1}{4}(1 - s)(+1) + \frac{1}{4}(1 + s)(+1) - \frac{1}{4}(1 + s)(-1) \right] \\
&+ b \left[-\frac{1}{4}(1 - s)(-1) + \frac{1}{4}(1 - s)(-1) + \frac{1}{4}(1 + s)(+1) - \frac{1}{4}(1 + s)(+1) \right] \\
&= \frac{a}{4} \left[(1 - s) + (1 - s) + (1 + s) + (1 + s) \right] \\
&+ \frac{b}{4} \left[(1 - s) - (1 - s) + (1 + s) - (1 + s) \right] \\
&= a
\end{aligned}
\tag{8.37}
$$

Here again, we find that the derivative of the bilinear field inside the element is exact: $\frac{\partial v^h}{\partial r} = \frac{\partial v}{\partial r}$.

However, following the same methodology as above, one can easily prove that this is no more true for polynomials of degree strictly higher than 1. This fact has serious consequences: if the solution to the problem at hand is for instance a parabola, the Q_1 basis functions cannot represent the solution properly, but only by approximating the parabola in each element by a line. As we will see later, Q_2 basis functions can remedy this problem by containing quadratic terms.

Remark. *The Q_1 basis functions are first-order polynomials. We have seen that they can be used to compute gradients. However they cannot be used to compute 2nd-order derivatives since their 2nd-order derivative is identically zero.*

An identical approach to arrive at the basis functions is presented in Ergatoudis, Irons, and Zienkiewicz [402] (1968).

8.3.2 Biquadratic basis functions (Q_2)

This element is part of the so-called Lagrange family [1114]. Inside an element the local numbering of the nodes is as follows[6]:
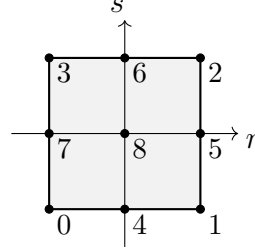
Note that this numbering is also employed in Li [825, p56]. The polynomial representation of the function ϕ over this element is then taken to be biquadratic:

$$\phi^h(r,s) = a + br + cs + drs + er^2 + fs^2 + gr^2s + hrs^2 + ir^2s^2 = \sum_{i=0}^{8} \mathcal{N}_i(r,s)\phi_i$$

and one can show that the basis functions are:

$$
\begin{aligned}
\mathcal{N}_0(r,s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \\
\mathcal{N}_1(r,s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \\
\mathcal{N}_2(r,s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s+1) \\
\mathcal{N}_3(r,s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \\
\mathcal{N}_4(r,s) &= (1-r^2)\frac{1}{2}s(s-1) \\
\mathcal{N}_5(r,s) &= \frac{1}{2}r(r+1)(1-s^2) \\
\mathcal{N}_6(r,s) &= (1-r^2)\frac{1}{2}s(s+1) \\
\mathcal{N}_7(r,s) &= \frac{1}{2}r(r-1)(1-s^2) \\
\mathcal{N}_8(r,s) &= (1-r^2)(1-s^2)
\end{aligned}
$$

Note that we have $\mathcal{N}_i(r_j,s_j) = \delta_{ij}$ and then obviously $\mathcal{N}_i(r_i,s_i) = 1$.

---
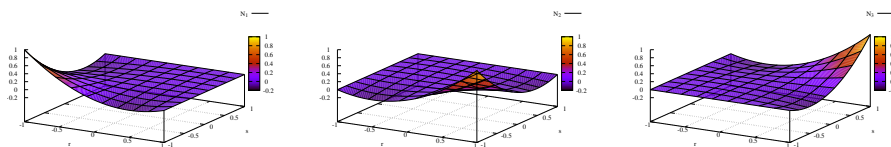
[6]I have adopted here a numbering scheme starting at zero! Also, it is a numbering among many other possible choices!

Surface representation of the basis functions on the reference element. in images/basis_Q2_2D/

Their derivatives are given by:

$$\frac{\partial \mathcal{N}_0}{\partial r} = \frac{1}{2}(2r-1)\frac{1}{2}s(s-1)$$

$$\frac{\partial \mathcal{N}_1}{\partial r} = \frac{1}{2}(2r+1)\frac{1}{2}s(s-1)$$

$$\frac{\partial \mathcal{N}_2}{\partial r} = \frac{1}{2}(2r+1)\frac{1}{2}s(s+1)$$

$$\frac{\partial \mathcal{N}_3}{\partial r} = \frac{1}{2}(2r-1)\frac{1}{2}s(s+1)$$

$$\frac{\partial \mathcal{N}_4}{\partial r} = (-2r)\frac{1}{2}s(s-1)$$

$$\frac{\partial \mathcal{N}_5}{\partial r} = \frac{1}{2}(2r+1)(1-s^2)$$

$$\frac{\partial \mathcal{N}_6}{\partial r} = (-2r)\frac{1}{2}s(s+1)$$

$$\frac{\partial \mathcal{N}_7}{\partial r} = \frac{1}{2}(2r-1)(1-s^2)$$

$$\frac{\partial \mathcal{N}_8}{\partial r} = (-2r)(1-s^2)$$

$$\frac{\partial \mathcal{N}_0}{\partial s} = \frac{1}{2}r(r-1)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_1}{\partial s} = \frac{1}{2}r(r+1)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_2}{\partial s} = \frac{1}{2}r(r+1)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_3}{\partial s} = \frac{1}{2}r(r-1)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_4}{\partial s} = (1-r^2)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_5}{\partial s} = \frac{1}{2}r(r+1)(-2s)$$

$$\frac{\partial \mathcal{N}_6}{\partial s} = (1-r^2)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_7}{\partial s} = \frac{1}{2}r(r-1)(-2s)$$

$$\frac{\partial \mathcal{N}_8}{\partial s} = (1-r^2)(-2s)$$

These basis functions are used for example in STONE 18.

8.3.3 Bicubic basis functions (Q_3)

basis_Q3_2D.tex

Inside an element a possible local numbering of the nodes is as follows:

(tikz_q32d.tex)

282
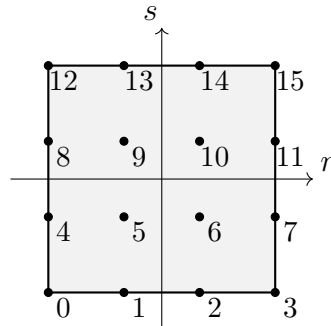
As shown in Section 8.2.3 the 1D cubic basis functions are given by:

$$\mathcal{N}_1(r) = (-1 + r + 9r^2 - 9r^3)/16 \qquad \mathcal{N}_1(s) = (-1 + s + 9s^2 - 9s^3)/16$$
$$\mathcal{N}_2(r) = (+9 - 27r - 9r^2 + 27r^3)/16 \qquad \mathcal{N}_2(s) = (+9 - 27s - 9s^2 + 27s^3)/16$$
$$\mathcal{N}_3(r) = (+9 + 27r - 9r^2 - 27r^3)/16 \qquad \mathcal{N}_3(s) = (+9 + 27s - 9s^2 - 27s^3)/16$$
$$\mathcal{N}_4(r) = (-1 - r + 9r^2 + 9r^3)/16 \qquad \mathcal{N}_4(s) = (-1 - s + 9s^2 + 9s^3)/16$$

and the resulting 2D basis functions are simply the tensor product of the above 1D ones:

$$
\begin{aligned}
\mathcal{N}_{01}(r,s) &= \mathcal{N}_1(r)\mathcal{N}_1(s) = (-1 + r + 9r^2 - 9r^3)/16 \cdot (-1 + s + 9s^2 - 9s^3)/16 \\
\mathcal{N}_{02}(r,s) &= \mathcal{N}_2(r)\mathcal{N}_1(s) = (+9 - 27r - 9r^2 + 27r^3)/16 \cdot (-1 + s + 9s^2 - 9s^3)/16 \\
\mathcal{N}_{03}(r,s) &= \mathcal{N}_3(r)\mathcal{N}_1(s) = (+9 + 27r - 9r^2 - 27r^3)/16 \cdot (-1 + s + 9s^2 - 9s^3)/16 \\
\mathcal{N}_{04}(r,s) &= \mathcal{N}_4(r)\mathcal{N}_1(s) = (-1 - r + 9r^2 + 9r^3)/16 \cdot (-1 + s + 9s^2 - 9s^3)/16 \\
\mathcal{N}_{05}(r,s) &= \mathcal{N}_1(r)\mathcal{N}_2(s) = (-1 + r + 9r^2 - 9r^3)/16 \cdot (9 - 27s - 9s^2 + 27s^3)/16 \\
\mathcal{N}_{06}(r,s) &= \mathcal{N}_2(r)\mathcal{N}_2(s) = (+9 - 27r - 9r^2 + 27r^3)/16 \cdot (9 - 27s - 9s^2 + 27s^3)/16 \\
\mathcal{N}_{07}(r,s) &= \mathcal{N}_3(r)\mathcal{N}_2(s) = (+9 + 27r - 9r^2 - 27r^3)/16 \cdot (9 - 27s - 9s^2 + 27s^3)/16 \\
\mathcal{N}_{08}(r,s) &= \mathcal{N}_4(r)\mathcal{N}_2(s) = (-1 - r + 9r^2 + 9r^3)/16 \cdot (9 - 27s - 9s^2 + 27s^3)/16 \\
\mathcal{N}_{09}(r,s) &= \mathcal{N}_1(r)\mathcal{N}_3(s) = (-1 + r + 9r^2 - 9r^3)/16 \cdot (9 + 27s - 9s^2 - 27s^3)/16 \\
\mathcal{N}_{10}(r,s) &= \mathcal{N}_2(r)\mathcal{N}_3(s) = (+9 - 27r - 9r^2 + 27r^3)/16 \cdot (9 + 27s - 9s^2 - 27s^3)/16 \\
\mathcal{N}_{11}(r,s) &= \mathcal{N}_3(r)\mathcal{N}_3(s) = (+9 + 27r - 9r^2 - 27r^3)/16 \cdot (9 + 27s - 9s^2 - 27s^3)/16 \\
\mathcal{N}_{12}(r,s) &= \mathcal{N}_4(r)\mathcal{N}_3(s) = (-1 - r + 9r^2 + 9r^3)/16 \cdot (9 + 27s - 9s^2 - 27s^3)/16 \\
\mathcal{N}_{13}(r,s) &= \mathcal{N}_1(r)\mathcal{N}_4(s) = (-1 + r + 9r^2 - 9r^3)/16 \cdot (-1 - s + 9s^2 + 9s^3)/16 \\
\mathcal{N}_{14}(r,s) &= \mathcal{N}_2(r)\mathcal{N}_4(s) = (+9 - 27r - 9r^2 + 27r^3)/16 \cdot (-1 - s + 9s^2 + 9s^3)/16 \\
\mathcal{N}_{15}(r,s) &= \mathcal{N}_3(r)\mathcal{N}_4(s) = (+9 + 27r - 9r^2 - 27r^3)/16 \cdot (-1 - s + 9s^2 + 9s^3)/16 \\
\mathcal{N}_{16}(r,s) &= \mathcal{N}_4(r)\mathcal{N}_4(s) = (-1 - r + 9r^2 + 9r^3)/16 \cdot (-1 - s + 9s^2 + 9s^3)/16 \quad (8.38)
\end{aligned}
$$
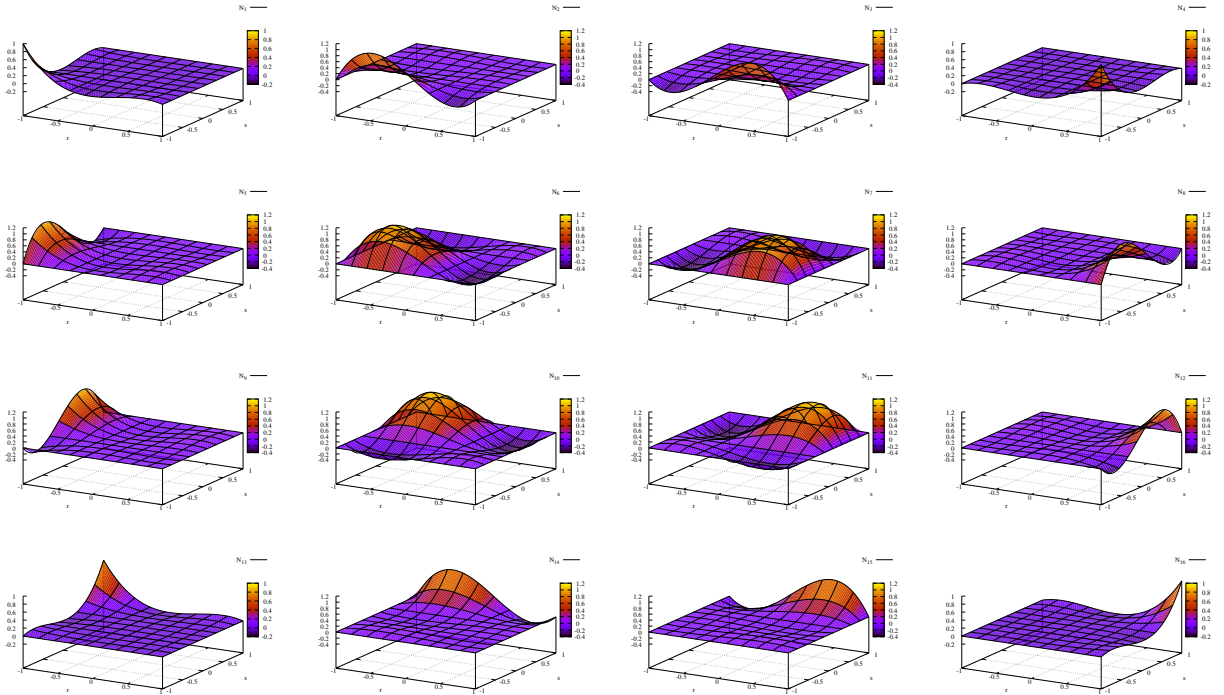
Surface representation of the basis functions on the reference element. in images/basis_Q3_2D/

The derivatives are trivial to obtain from the derivatives of the 1D basis functions, e.g.

$$\frac{\partial \mathcal{N}_{13}}{\partial r} = \frac{\partial \mathcal{N}_1}{\partial r} \mathcal{N}_3(s)$$

These basis functions are used in STONE 19.

8.3.4 Bicubic basis functions in 2D ($Q_3^{(12)}$)

This element is (as far as I know) not widely used. I have found it in Ergatoudis, Irons, and Zienkiewicz [402] (1968) - check the appendix of the paper too. In essence the internal nodes of the Q_3 are absent, which leaves 12 nodes (instead of 16).
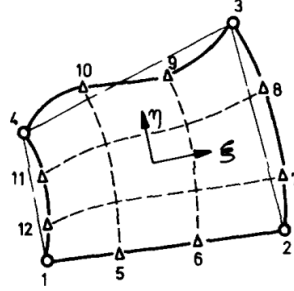
8.3.5 Eight node serendipity basis functions in 2D ($Q_2^{(8)}$)

The serendipity elements are those rectangular elements which have no interior nodes (See for example Reddy [1118, p65]). Inside an element a possible local numbering of the nodes is as follows:
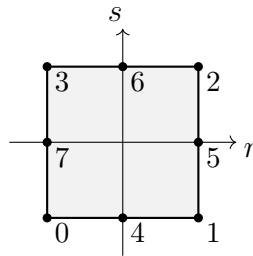
The main difference with the Q_2 element resides in the fact that there is no node in the middle of the element. The polynomial representation of the function ϕ over the element is then

$$\phi_h(r, s) = a + br + cs + drs + er^2 + fs^2 + gr^2s + hrs^2$$

Note that absence of the r^2s^2 term which was previously associated to the center node. We find that

$$\mathcal{N}_0(r,s) = \frac{1}{4}(1-r)(1-s)(-r-s-1) \tag{8.39}$$

$$\mathcal{N}_1(r,s) = \frac{1}{4}(1+r)(1-s)(r-s-1) \tag{8.40}$$

$$\mathcal{N}_2(r,s) = \frac{1}{4}(1+r)(1+s)(r+s-1) \tag{8.41}$$

$$\mathcal{N}_3(r,s) = \frac{1}{4}(1-r)(1+s)(-r+s-1) \tag{8.42}$$

$$\mathcal{N}_4(r,s) = \frac{1}{2}(1-r^2)(1-s) \tag{8.43}$$

$$\mathcal{N}_5(r,s) = \frac{1}{2}(1+r)(1-s^2) \tag{8.44}$$

$$\mathcal{N}_6(r,s) = \frac{1}{2}(1-r^2)(1+s) \tag{8.45}$$

$$\mathcal{N}_7(r,s) = \frac{1}{2}(1-r)(1-s^2) \tag{8.46}$$

The basis functions at the mid side nodes are products of a second order polynomial parallel to side and a linear function perpendicular to the side while basis functions for corner nodes are modifications of the bilinear quadrilateral element.
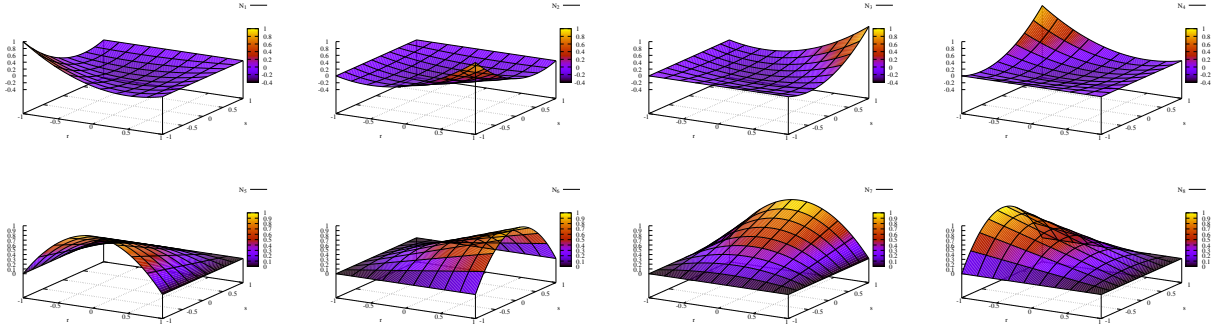
Surface representation of the basis functions on the reference element. in images/basis_Q28_2D/

The first-order derivatives are given by:

$$\frac{\partial \mathcal{N}_0}{\partial r}(r,s) = -\frac{1}{4}(s-1)(2r+s) \tag{8.47}$$

$$\frac{\partial \mathcal{N}_1}{\partial r}(r,s) = -\frac{1}{4}(s-1)(2r-s) \tag{8.48}$$

$$\frac{\partial \mathcal{N}_2}{\partial r}(r,s) = \frac{1}{4}(s+1)(2r+s) \tag{8.49}$$

$$\frac{\partial \mathcal{N}_3}{\partial r}(r,s) = \frac{1}{4}(s+1)(2r-s) \tag{8.50}$$

$$\frac{\partial \mathcal{N}_4}{\partial r}(r,s) = r(s-1) \tag{8.51}$$

$$\frac{\partial \mathcal{N}_5}{\partial r}(r,s) = \frac{1}{2}(1-s^2) \tag{8.52}$$

$$\frac{\partial \mathcal{N}_6}{\partial r}(r,s) = -r(s+1) \tag{8.53}$$

$$\frac{\partial \mathcal{N}_7}{\partial r}(r,s) = -\frac{1}{2}(1-s^2) \tag{8.54}$$

$$\frac{\partial \mathcal{N}_0}{\partial s}(r,s) = -\frac{1}{4}(r-1)(r+2s) \tag{8.55}$$

$$\frac{\partial \mathcal{N}_1}{\partial s}(r,s) = -\frac{1}{4}(r+1)(r-2s) \tag{8.56}$$

$$\frac{\partial \mathcal{N}_2}{\partial s}(r,s) = \frac{1}{4}(r+1)(r+2s) \tag{8.57}$$

$$\frac{\partial \mathcal{N}_3}{\partial s}(r,s) = \frac{1}{4}(r-1)(r-2s) \tag{8.58}$$

$$\frac{\partial \mathcal{N}_4}{\partial s}(r,s) = -\frac{1}{2}(1-r^2) \tag{8.59}$$

$$\frac{\partial \mathcal{N}_5}{\partial s}(r,s) = -(r+1)s \tag{8.60}$$

$$\frac{\partial \mathcal{N}_6}{\partial s}(r,s) = \frac{1}{2}(1-r^2) \tag{8.61}$$

$$\frac{\partial \mathcal{N}_7}{\partial s}(r,s) = (r-1)s \tag{8.62}$$

These basis functions are used in STONE 52. An identical approach to arrive at the basis functions is presented in Ergatoudis, Irons, and Zienkiewicz [402] (1968).

8.3.6 Eight node serendipity basis functions in 2D ($QH8 - C1$)

This element is proposed in Zhang and Xiang [1505] (2020). Two remarks must be made: 1) Eq. (29) of their publication which is the definition of the basis functions contains an error[7]. 2) The authors use a rather uncommon and annoying rotated numbering:

y

---

[7]Answer from the author: "N5 to N8 is missing an A in the denominator and the calculation program does not have this problem"

```
      |
2=====5=====1            3=====6=====2
|           |            |           |    (r_0,s_0)=(-1,-1)    (r_4,s_4)=( 0,-1)
|           |            |           |    (r_1,s_1)=(+1,-1)    (r_5,s_5)=(+1, 0)
6           8--x         7     +     5    (r_2,s_2)=(+1,+1)    (r_6,s_6)=( 0,+1)
|           |            |           |    (r_3,s_3)=(-1,+1)    (r_7,s_7)=(-1, 0)
|           |            |           |
3=====7=====4            0=====4=====1
Zhang & Xiang            our numbering
```

For each element they define (their numbering):

$$A = \frac{1}{2}[(x_1 - x_3)(y_2 - y_4) - (x_2 - x_4)(y_1 - y_3)]$$
$$m_x = (x_1 - x_4)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_4)$$
$$m_y = (x_3 - x_4)(y_1 - y_2) - (x_1 - x_2)(y_3 - y_4)$$

Note that A is the area of the element, and that in the case when the element is a rectangle then $m_x = m_y = 0$.

$$\mathcal{N}_1(r, s) = n_1(r, s) + (m_x^2 - m_x m_y + m_y^2)\frac{E(r, s)}{D}$$
$$\mathcal{N}_2(r, s) = n_2(r, s) + (m_x^2 + m_x m_y + m_y^2)\frac{E(r, s)}{D}$$
$$\mathcal{N}_3(r, s) = n_3(r, s) + (m_x^2 - m_x m_y + m_y^2)\frac{E(r, s)}{D}$$
$$\mathcal{N}_4(r, s) = n_4(r, s) + (m_x^2 + m_x m_y + m_y^2)\frac{E(r, s)}{D}$$
$$\mathcal{N}_5(r, s) = n_5(r, s) - m_x(2Am_x + m_y^2)\frac{E(r, s)}{AD}$$
$$\mathcal{N}_6(r, s) = n_6(r, s) - m_y(2Am_y + m_x^2)\frac{E(r, s)}{AD}$$
$$\mathcal{N}_7(r, s) = n_7(r, s) + m_x(-2Am_x + m_y^2)\frac{E(r, s)}{AD}$$
$$\mathcal{N}_8(r, s) = n_8(r, s) + m_y(-2Am_y + m_x^2)\frac{E(r, s)}{AD}$$

with

$$E(r, s) = (1 - r^2)(1 - s^2) \qquad D = 4(4A^2 + m_x^2 + m_y^2)$$

and where the n_i functions are the basis functions of the 'regular' 8-node element (see Section 8.3.5).
This is implemented in STONE 52.

not finished. SHOW CONSISTENCY !! like in paper email sent to author about mistake.

Let us verify consistency:

$$\sum_{i=1}^{8} \mathcal{N}_i(r, s) = \underbrace{\sum_{i=1}^{8} n_i(r, s)}_{=0} + \frac{E(r, s)}{D}\left[(m_x^2 - m_x m_y + m_y^2) + (m_x^2 + m_x m_y + m_y^2) + (m_x^2 - m_x m_y + m_y^2) + (m\right.$$

$$\left. - m_x(2Am_x + m_y^2)\frac{1}{A} - m_y(2Am_y + m_x^2)\frac{1}{A} + m_x(-2Am_x + m_y^2)\frac{1}{A} + m_y(-2Am_y + m_x^2)\frac{1}{A}\right]$$

$$= \frac{E(r, s)}{D}\left[(4m_x^2 + 4m_y^2) + \frac{1}{A}(-4Am_x^2 - 4Am_y^2)\right]$$

$$= 0$$

287

8.3.7 Biquartic basis functions in 2D (Q_4)

Inside an element the local numbering of the nodes is as follows:
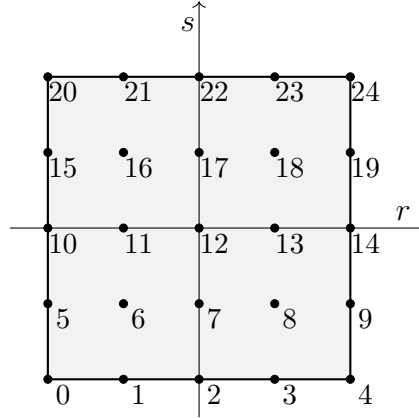
8.3.8 Enriched linear basis functions in quadrilaterals (Q_1^+) -WIP

```
4===========3
|            |    (r_1,s_1)=(-1,-1)
|            |    (r_2,s_2)=(1,-1)
|      5     |    (r_3,s_3)=(1,1)
|            |    (r_4,s_4)=(-1,1)
|            |    (r_5,s_5)=(0,0)
1===========2
```

- In Bai [42] (1997): "It is well known that the equal-order bilinear velocity-bilinear continuous pressure element - the $Q_1 \times Q_1$, element - exhibits a certain spurious pressure mode. In the paper we propose a new stabilized $Q_1 \times Q_1$ combination for the velocity and pressure with three internal degrees of freedom added to the velocity space, that is, one degree of freedom for each component of the velocity and one degree of freedom shared by both components of the velocity."

Two versions are proposed, if I understand it correctly. The first one is given in Eq. (7) (three extra dofs: u_5, v_5, w):

$$
\begin{aligned}
u^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)u_i + \left[u_5 - \frac{w}{4}(1-s) \right](1-r^2)(1-s^2) \\
v^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)v_i + \left[v_5 - \frac{w}{4}(1-r) \right](1-r^2)(1-s^2)
\end{aligned}
\tag{8.65}
$$

The second one in Eq. (23) (four extra dofs: u_5, v_5, u_6, v_6):

$$
\begin{aligned}
u^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)u_i + \left[u_5 + u_6(r+s) \right](1-r^2)(1-s^2) \\
v^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)v_i + \left[v_5 + v_6(r+s) \right](1-r^2)(1-s^2)
\end{aligned}
\tag{8.66}
$$

- In Franca, Oliveira, and Sarkis [437] (2007): "Stabilized finite element method for Stokes equations with piecewise continuous bilinear approximations for both velocity and pressure variables. The velocity field is enriched with piecewise polynomial bubble functions with null average at element edges."

It looks like they are proposing (see their Eq. (2.6)):

$$
\begin{aligned}
u^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)u_i + (\alpha + \gamma s)\frac{1}{2}(r^2 + s^2 - \frac{4}{3}) \\
v^h(r,s) &= \sum_{i=1}^{4} N_i(r,s)v_i + (\beta + \gamma r)\frac{1}{2}(r^2 + s^2 - \frac{4}{3})
\end{aligned} \tag{8.67}
$$

- In Kwon and Park [782] (2014): "We introduce a new stable MINI-element pair for incompressible Stokes equations on quadrilateral meshes, which uses the smallest number of bubbles for the velocity. The pressure is discretized with the P_1-midpoint-edge-continuous elements and each component of the velocity field is done with the standard Q_1-conforming elements enriched by one bubble a quadrilateral."

- In Lamichhane [784] (2017): "We consider a quadrilateral MINI finite element for approximating the solution of Stokes equations using a quadrilateral mesh. We use the standard bilinear finite element space enriched with element-wise defined bubble functions for the velocity and the standard bilinear finite element space for the pressure space. With a simple modification of the standard bubble function we show that a single bubble function is sufficient to ensure the inf-sup condition. This is a refinement of Bai (1997) [42] where the author enriches the velocity space with more than a single vector bubble function per element. In this article we show that with a small modification of the standard bubble function we can get the stability just by using a single vector bubble function per element."

lamichhane2D.tex

The two bubble functions are defined on the reference element $[-1,1] \times [-1,1]$:

$$
\begin{aligned}
b^{(1)}(r,s) &= (1-r)(1-s) \cdot (1-r^2)(1-s^2) \tag{8.68} \\
b^{(2)}(r,s) &= \left(1 + \frac{r+s}{4}\right) \cdot (1-r^2)(1-s^2) \tag{8.69}
\end{aligned}
$$

Both bubble functions are exactly one in the middle of the element and exactly zero on the edges of the element as expected from basis functions.

We then have

$$\frac{\partial b^{(1)}}{\partial r}(r,s) = (1-s)^2(1+s)[-2(1-r)(1+r) + (1-r)^2]$$
$$= (1-s)^2(1+s)[-2 + 2r^2 + 1 - 2r + r^2]$$
$$= (1-s)^2(1+s)[-1 - 2r + 3r^2] \tag{8.70}$$

$$\frac{\partial b^{(1)}}{\partial s}(r,s) = (1-r)^2(1+r)[-1 - 2s + 3s^2] \tag{8.71}$$

$$\frac{\partial b^{(2)}}{\partial r}(r,s) = \frac{1}{4}(1-s^2)(1 - r^2 + (4+r+s)(-2r))$$
$$= \frac{1}{4}(1-s^2)(1 - 8r - 3r^2 - 2rs) \tag{8.72}$$

$$\frac{\partial b^{(2)}}{\partial s}(r,s) = \frac{1}{4}(1-r^2)(1 - s^2 + (4+r+s)(-2s))$$
$$= \frac{1}{4}(1-r^2)(1 - 8s - 3s^2 - 2rs) \tag{8.73}$$

We postulate that a function f has the following representation in the element:

$$f^h(r,s) = a + br + cs + drs + e\,b(r,s)$$

where $b(r,s)$ stands for the bubble function which is of the form $b(r,s) = (1-r^2)(1-s^2)\phi(r,s)$ and ϕ is a (bi)-linear function of r, s.

We need

$$f^h(r_1, s_1) = a - b - c + d = f_1 \tag{8.74}$$
$$f^h(r_2, s_2) = a + b - c - d = f_2 \tag{8.75}$$
$$f^h(r_3, s_3) = a + b + c + d = f_3 \tag{8.76}$$
$$f^h(r_4, s_4) = a - b + c - d = f_4 \tag{8.77}$$
$$f^h(r_5, s_5) = a + e = f_5 \tag{8.78}$$

This can be written as a linear system:

$$\begin{pmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix}$$

and the solution is then:

$$\begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ -1 & 1 & 1 & -1 & 0 \\ -1 & -1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ -1 & -1 & -1 & -1 & 4 \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix}$$

290

or,

$$a = \frac{1}{4}(f_1 + f_2 + f_3 + f_4)$$

$$b = \frac{1}{4}(-f_1 + f_2 + f_3 - f_4)$$

$$c = \frac{1}{4}(-f_1 - f_2 + f_3 + f_4)$$

$$d = \frac{1}{4}(f_1 - f_2 + f_3 - f_4)$$

$$e = \frac{1}{4}(-f_1 - f_2 - f_3 - f_4 + 4f_5) \tag{8.79}$$

Then

$$
\begin{aligned}
4f^h(r,s) &= 4[a + br + cs + drs + e(1 - r^2)(1 - s^2)\phi(r,s)] \\
&= (f_1 + f_2 + f_3 + f_4) \\
&\quad + (-f_1 + f_2 + f_3 - f_4)r \\
&\quad + (-f_1 - f_2 + f_3 + f_4)s \\
&\quad + (f_1 - f_2 + f_3 - f_4)rs \\
&\quad + (-f_1 - f_2 - f_3 - f_4 + 4f_5)(1 - r^2)(1 - s^2)\phi(r,s) \\
&= (1 - r - s + rs - b(r,s))f_1 \\
&\quad + (1 + r - s - rs - b(r,s))f_2 \\
&\quad + (1 + r + s + rs - b(r,s))f_3 \\
&\quad + (1 - r + s - rs - b(r,s))f_4 \\
&\quad + 4b(r,s)f_5
\end{aligned}
\tag{8.80}
$$

or,

$$
\begin{aligned}
f^h(r,s) &= \underbrace{\left(\frac{1}{4}(1-r)(1-s) - \frac{1}{4}b(r,s)\right)}_{\mathcal{N}_1} f_1 + \underbrace{\left(\frac{1}{4}(1+r)(1-s) - \frac{1}{4}b(r,s)\right)}_{\mathcal{N}_2} f_2 \\
&\quad + \underbrace{\left(\frac{1}{4}(1+r)(1+s) - \frac{1}{4}b(r,s)\right)}_{\mathcal{N}_3} f_3 + \underbrace{\left(\frac{1}{4}(1-r)(1+s) - \frac{1}{4}b(r,s)\right)}_{\mathcal{N}_4} f_4 \\
&\quad + \underbrace{b(r,s)}_{\mathcal{N}_5} f_5
\end{aligned}
\tag{8.81}
$$

As in the P_1^+ case the resulting basis functions are a combination of the regular Q_1 basis functions and the bubble.

  – Zeroth-order consistency check $f(r,s) = C$:

$$f^h(r,s) = \sum_{i=1}^{5} \mathcal{N}_i(r,s)f_i = C \sum_{i=1}^{5} \mathcal{N}_i(r,s) = C \tag{8.82}$$

291

- First-order consistency check $f(r, s) = r$ (or $f(r, s) = s$):

$$
\begin{aligned}
f^h(r, s) &= \sum_{i=1}^{5} \mathcal{N}_i(r, s) f_i \\
&= \mathcal{N}_1(r, s)(-1) + \mathcal{N}_2(r, s)(+1) + \mathcal{N}_3(r, s)(+1) + \mathcal{N}_4(r, s)(-1) + \mathcal{N}_5(r, s)(0) \\
&= -\mathcal{N}_1(r, s) + \mathcal{N}_2(r, s) + \mathcal{N}_3(r, s) - \mathcal{N}_4(r, s) \\
&= r
\end{aligned}
\tag{8.83}
$$

- Second-order consistency check $f(r, s) = rs$ ($f_1 = (-1)(-1) = 1$, $f_2 = (+1)(-1) = -1$, etc ...)

$$
\begin{aligned}
f^h(r, s) &= \sum_{i=1}^{5} \mathcal{N}_i(r, s) f_i \\
&= \mathcal{N}_1(r, s)(+1) + \mathcal{N}_2(r, s)(-1) + \mathcal{N}_3(r, s)(+1) + \mathcal{N}_4(r, s)(-1) + \mathcal{N}_5(r, s)(0) \\
&= \mathcal{N}_1 - \mathcal{N}_2 + \mathcal{N}_3 - \mathcal{N}_4 \\
&= \left(\frac{1}{4}(1-r)(1-s) - \frac{1}{4}b(r, s) \right) - \left(\frac{1}{4}(1+r)(1-s) - \frac{1}{4}b(r, s) \right) \\
&\quad + \left(\frac{1}{4}(1+r)(1+s) - \frac{1}{4}b(r, s) \right) - \left(\frac{1}{4}(1-r)(1+s) - \frac{1}{4}b(r, s) \right) \\
&= \frac{1}{4}(1-r)(1-s) - \frac{1}{4}(1+r)(1-s) + \frac{1}{4}(1+r)(1+s) - \frac{1}{4}(1-r)(1+s) \\
&= \frac{1}{2}(-r)(1-s) + \frac{1}{2}(+r)(1+s) \\
&= rs
\end{aligned}
\tag{8.84}
$$

We find that the basis functions can represent a bilinear field exactly.

Consistency check for quadratic terms, i.e. $f(r, s) = r^2$ (or $f(r, s) = s^2$):

$$
\begin{aligned}
f^h(r, s) &= \sum_{i=1}^{5} \mathcal{N}_i(r, s) f_i \\
&= \mathcal{N}_1(r, s) \cdot (+1) + \mathcal{N}_2(r, s) \cdot (+1) + \mathcal{N}_3(r, s) \cdot (+1) + \mathcal{N}_4(r, s) \cdot (+1) + \mathcal{N}_5(r, s) \cdot (0) \\
&= \left(\frac{1}{4}(1-r)(1-s) - \frac{1}{4}b(r, s) \right) + \left(\frac{1}{4}(1+r)(1-s) - \frac{1}{4}b(r, s) \right) \\
&\quad + \left(\frac{1}{4}(1+r)(1+s) - \frac{1}{4}b(r, s) \right) + \left(\frac{1}{4}(1-r)(1+s) - \frac{1}{4}b(r, s) \right) \\
&= \frac{1}{2}(1-s) + \frac{1}{2}(1+s) - b(r, s) \\
&= 1 - b(r, s)
\end{aligned}
\tag{8.85}
$$

We have

$$
\int_{-1}^{+1} \int_{-1}^{+1} (1 - b_1(r, s)) dr ds = \int_{-1}^{+1} \int_{-1}^{+1} [1 - (1 - r^2)(1 - s^2)(1 - r)(1 - s)] dr ds = 20/9 \simeq 2.2\!
$$

$$
\int_{-1}^{+1} \int_{-1}^{+1} (1 - b_2(r, s, \beta)) dr ds = \int_{-1}^{+1} \int_{-1}^{+1} [1 - (1 - r^2)(1 - s^2)(1 + \beta(r + s))] dr ds = 20/9 \qquad \forall
$$

Both bubbles yield the same average. This is not helpful.

Let us now look at the (root) mean square:

$$\int_{-1}^{+1}\int_{-1}^{+1}(1-b_1(r,s))^2 drds = 21284/11025 \simeq 1.93052$$

$$\int_{-1}^{+1}\int_{-1}^{+1}(1-b_2(r,s,\beta))^2 drds = \frac{4}{1575}(128\beta^2+623) \quad\quad (8.86)$$

The problem is that the minimum is reached for $\beta = 0$ which is not allowed so we cannot choose β so as to minimise the error. For $\beta = 0.25$ as used in the paper:

$$\int_{-1}^{+1}\int_{-1}^{+1}(1-b_2(r,s))^2 drds = 2524/1575 \simeq 1.60254$$

On the other hand, this means that using the second bubble function does a better job at representing square terms (r^2, s^2) than using the first one.

One can also revisit the second bubble function: in Lamichhane (2017) [784] it is postulated to be defined by

$$b^{(2)}(r,s) = (a+br+cs)(1-r^2)(1-s^2) \quad\quad abc \neq 0 \quad\quad (8.87)$$

on the reference element $[-1,1]\times[-1,1]$. Then the author states that 'for simplicity we choose':

$$b^{(2)}(r,s) = \frac{1}{4}(4+r+s)(1-r^2)(1-s^2) \quad\quad (8.88)$$

and that 'the factor $1/4$ is used to force the value of the bubble function at the centroid of the square to be 1'.

Looking closer, we see that forcing the bubble to be 1 in $(r,s) = (0,0)$ does impose $a = 1$ but leaves b, c free, i.e. the bubble is then:

$$b^{(2)}(r,s) = (1+br+cs)(1-r^2)(1-s^2) \quad\quad bc \neq 0 \quad\quad (8.89)$$

For symmetry reasons I would be tempted to indeed take $b = c$ but I am then left with

$$b^{(2)}(r,s) = [1+b(r+s)](1-r^2)(1-s^2) \quad\quad b \neq 0 \quad\quad (8.90)$$

which means that Lamichhane sets $b = c = 1/4$ in his paper.

Question: We know that $b = 0$ is not allowed, but could it not be possible to design an analytical or numerical test or a theory to choose an 'optimal' value (in some sense) for b?

_____ -

Let us consider a square mesh with $nelx^2$ elements for simplicity. The number of V dofs for a Q_1 space would be $(nelx+1)^2 = nelx^2 + 2nelx + 1$. The number of V dofs for a Q_1^+ space would be $(nelx+1)^2 + nelx^2 = 2nelx^2 + 2nelx + 1$. The number of V dofs for a Q_2 space would be $(2*nelx+1)^2 = 4nelx^2 + 4nelx + 1$. Asymptotically, for large values of $nelx$, we find that a Q_1^+ space requires twice as many dofs as Q_1 while Q_2 requires 4 times as many.
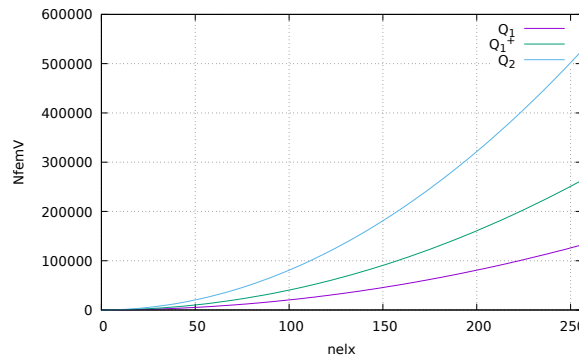
📚 Relevant Literature:

- Mons & Roge (1992) [944],

- Li *et al.* (2009) [827],

- Knobloch & Tobiska (2000) [757],

- Franca *et al.* (1993) [435],

- Idelsohn *et al.* (1995) [651].

8.3.9 The rotated Q_1 (Rannacher-Turek element)

basis_q1rc_2D.tex

The nodes are not on the corners of the element but in the middle of the element edges:

(tikz_RTQ1P0.tex)


```
+======3======+
|             |
|      s      |
|      |      |
4      +--r   2
|             |
|             |
|             |
+======1======+
```
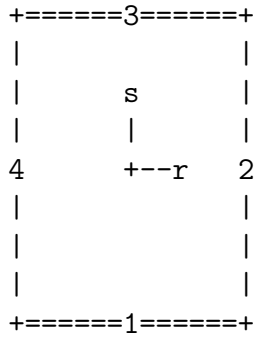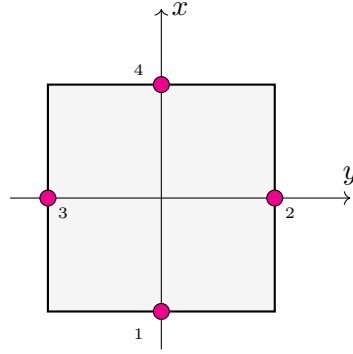
There are two types of basis functions: the Middle Point (MP) variant such that $\mathcal{N}_i(\vec{r}_j) = \delta_{ij}$ and the Mid Value (MV) variant such that $\frac{1}{|\Gamma_i|} \int_{\Gamma_i} \mathcal{N}_j d\Gamma = \delta_{ij}$.

The Middle Point (MP) variant . We have $\tilde{Q}_1 = span\{1, r, s, r^2 - s^2\}$ so a function $f \in \tilde{Q}_1$ is such that

$$f(r,s) = a + br + cs + d(r^2 - s^2) \tag{8.91}$$

This function must be so that

$$
\begin{aligned}
f_1 &= f(r = 0, s = -1) = a - c - d & (8.92)\\
f_2 &= f(r = +1, s = 0) = a + b + d & (8.93)\\
f_3 &= f(r = 0, s = +1) = a + c - d & (8.94)\\
f_4 &= f(r = -1, s = 0) = a - b + d & (8.95)
\end{aligned}
$$

294

and then

$$
\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}
$$

This system can easily be solved, and a, b, c, d are then replaced in Eq. (8.91), which yields

$$
f(r, s) = \mathcal{N}_1(r, s) f_1 + \mathcal{N}_2(r, s) f_2 + \mathcal{N}_3(r, s) f_3 + \mathcal{N}_4(r, s) f_4 \tag{8.96}
$$

inside the element with

$$
\begin{aligned}
\mathcal{N}_1(r, s) &= \frac{1}{4}(1 - 2s - (r^2 - s^2)) \\
\mathcal{N}_2(r, s) &= \frac{1}{4}(1 + 2r + (r^2 - s^2)) \\
\mathcal{N}_3(r, s) &= \frac{1}{4}(1 + 2s - (r^2 - s^2)) \\
\mathcal{N}_4(r, s) &= \frac{1}{4}(1 - 2r + (r^2 - s^2))
\end{aligned}
$$

We of course recover the partition of unity property, i.e. $\sum \mathcal{N}_i(r, s) = 1$ for any coordinate r, s inside the reference element.

Remark. *These basis functions have been independently proposed by Donea, Giuliani, Morgan, and Quartapelle [362] (1981). The authors prove herein that this element is checkerboard-free (although they do no show any example of simulation carried out with this element).*

$$
\frac{\partial \mathcal{N}_1}{\partial r} = \frac{1}{2}(-r) \tag{8.97}
$$

$$
\frac{\partial \mathcal{N}_2}{\partial r} = \frac{1}{2}(1 + r) \tag{8.98}
$$

$$
\frac{\partial \mathcal{N}_3}{\partial r} = \frac{1}{2}(-r) \tag{8.99}
$$

$$
\frac{\partial \mathcal{N}_4}{\partial r} = \frac{1}{2}(-1 + r) \tag{8.100}
$$

$$
\frac{\partial \mathcal{N}_1}{\partial s} = \frac{1}{2}(-1 + s) \tag{8.101}
$$

$$
\frac{\partial \mathcal{N}_2}{\partial s} = \frac{1}{2}(-s) \tag{8.102}
$$

$$
\frac{\partial \mathcal{N}_3}{\partial s} = \frac{1}{2}(1 + s) \tag{8.103}
$$

$$
\frac{\partial \mathcal{N}_4}{\partial s} = \frac{1}{2}(-s) \tag{8.104}
$$
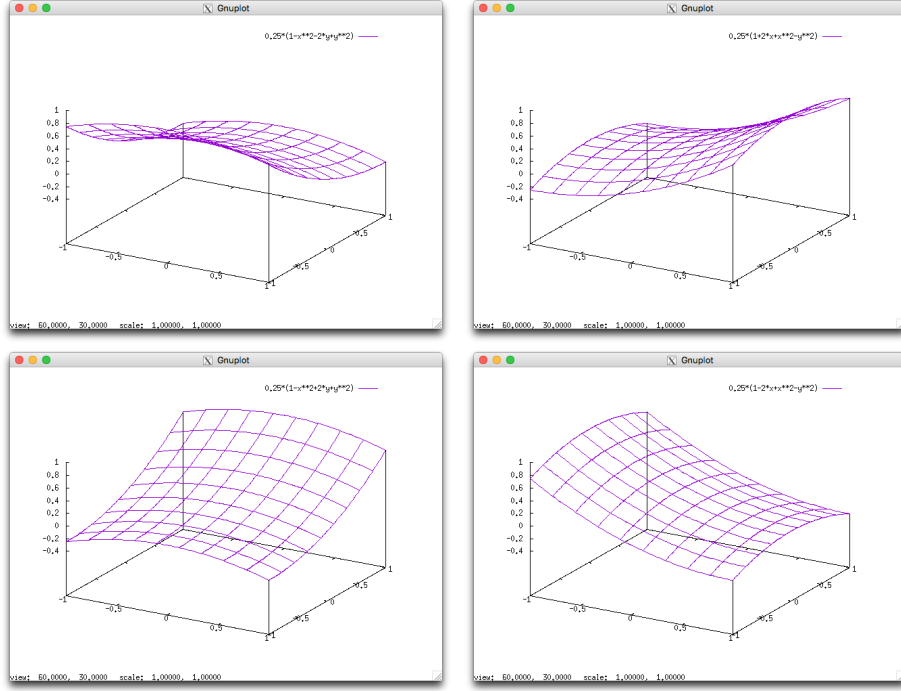
Graphical representation of the $\tilde{Q}_1$ basis functions

The Mid Value (MV) variant .

These basis functions are implemented in deal.II [8] for $x \in [0,1]$ and $y \in [0,1]$:

$$
\begin{align}
\mathcal{N}_1(x,y) &= 0.75 + 1.5x - 2.5y - 1.5(x^2 - y^2) \quad bottom \tag{8.105}\\
\mathcal{N}_2(x,y) &= -0.25 - 0.5x + 1.5y + 1.5(x^2 - y^2) \quad right \tag{8.106}\\
\mathcal{N}_3(x,y) &= -0.25 + 1.5x - 0.5y - 1.5(x^2 - y^2) \quad top \tag{8.107}\\
\mathcal{N}_4(x,y) &= 0.75 - 2.5x + 1.5y + 1.5(x^2 - y^2) \quad left \tag{8.108}
\end{align}
$$

We then proceed to rewrite these for $r \in [-1,1]$ and $t \in [-1:1]$:

$$
\begin{align}
\mathcal{N}_1(r,s) &= \frac{1}{4} - \frac{1}{2}s - \frac{3}{8}(r^2 - s^2) \quad bottom \tag{8.109}\\
\mathcal{N}_2(r,s) &= \frac{1}{4} + \frac{1}{2}r + \frac{3}{8}(r^2 - s^2) \quad right \tag{8.110}\\
\mathcal{N}_3(r,s) &= \frac{1}{4} + \frac{1}{2}s - \frac{3}{8}(r^2 - s^2) \quad top \tag{8.111}\\
\mathcal{N}_4(r,s) &= \frac{1}{4} - \frac{1}{2}r + \frac{3}{8}(r^2 - s^2) \quad left \tag{8.112}
\end{align}
$$

It is easy to verify that these functions verify the property

$$
\frac{1}{|\Gamma_i|} \int_{\Gamma_i} N_j d\Gamma = \delta_{ij}
$$

These basis functions are used in Shipeng & Zhongci (2006) [1240] and mentioned in John [688,

---

[8]https://www.dealii.org/8.5.0/doxygen/deal.II/polynomials_rannacher_turek_8cc_source.html

$$\frac{\partial N_1}{\partial r} = -\frac{3}{4}r$$

$$\frac{\partial N_2}{\partial r} = \frac{1}{2} + \frac{3}{4}r$$

$$\frac{\partial N_3}{\partial r} = -\frac{3}{4}r$$

$$\frac{\partial N_4}{\partial r} = -\frac{1}{2} + \frac{3}{4}r$$

$$\frac{\partial N_1}{\partial t} = -\frac{1}{2} + \frac{3}{4}t$$

$$\frac{\partial N_2}{\partial t} = -\frac{3}{4}t$$

$$\frac{\partial N_3}{\partial t} = \frac{1}{2} + \frac{3}{4}t$$

$$\frac{\partial N_4}{\partial t} = -\frac{3}{4}t$$

8.3.10 Linear basis functions for triangles in 2D (P_1)

basis_P1_2D.tex

Here we do not start from a reference element but consider instead a generic triangle:

(tikz_P1.tex)
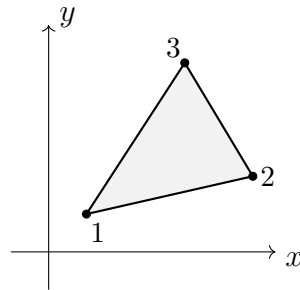
This is the simplest 2D element, which is also called linear triangular element. Velocities (or displacements) (u^h, v^h) in the element are interpolated from nodal velocities (u_i, v_i) using basis functions $\mathcal{N}_i$ as follows,

$$\vec{v}^h = \begin{pmatrix} u^h(x,y) \\ v^h(x,y) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^3 \mathcal{N}_i(x,y)u_i \\ \sum_{i=1}^3 \mathcal{N}_i(x,y)v_i \end{pmatrix} = \begin{pmatrix} \mathcal{N}_1(x,y) & 0 & \mathcal{N}_2(x,y) & 0 & \mathcal{N}_3(x,y) & 0 \\ 0 & \mathcal{N}_1(x,y) & 0 & \mathcal{N}_2(x,y) & 0 & \mathcal{N}_3(x,y) \end{pmatrix}$$

or simply

For this element, we have three nodes at the vertices of the triangle, which are numbered around the element in the counterclockwise direction. Each node has two degrees of freedom (i.e. it can

297

move in the x and y directions). The velocities u^h and v^h are assumed to be linear functions within the element, that is,

$$
\begin{aligned}
u^h(x,y) &= b_1 + b_2 x + b_3 y \\
v^h(x,y) &= b_4 + b_5 x + b_6 y
\end{aligned}
\tag{8.113}
$$

where b_i are constants to be determined and which depend on the triangle shape. Note that the strain rate components are then given by

$$
\begin{aligned}
\dot{\varepsilon}_{xx}(\vec{v}) &= b_2 \\
\dot{\varepsilon}_{yy}(\vec{v}) &= b_6 \\
\dot{\varepsilon}_{xy}(\vec{v}) &= (b_3 + b_5)/2
\end{aligned}
$$

and are constant throughout the element.

The velocities should satisfy the following six equations (when it is evaluated at a node we should recover the nodal velocity):

$$
\begin{aligned}
u_1 &= u^h(x_1, y_1) = b_1 + b_2 x_1 + b_3 y_1 \\
u_2 &= u^h(x_2, y_2) = b_1 + b_2 x_2 + b_3 y_2 \\
u_3 &= u^h(x_3, y_3) = b_1 + b_2 x_3 + b_3 y_3 \\
v_1 &= v^h(x_1, y_1) = b_4 + b_5 x_1 + b_6 y_1 \\
v_2 &= v^h(x_2, y_2) = b_4 + b_5 x_2 + b_6 y_2 \\
v_3 &= v^h(x_3, y_3) = b_4 + b_5 x_3 + b_6 y_3
\end{aligned}
$$

Let us focus on the three equations with the u component of the velocity. These can be re-written:

$$
\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}
$$

In order to obtain b_1, b_2, b_3 we need to solve this system, or simply to compute the inverse of the 3×3 M matrix, as explained in Appendix 3.1.5. We define $D = \det(M)$ and we get

$$
\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \frac{1}{D} \tilde{M} \cdot \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}
$$

The matrix $\tilde{M}$ is given by:

$$
\tilde{M} = \begin{pmatrix}
x_2 y_3 - x_3 y_2 & x_3 y_1 - x_1 y_3 & x_1 y_2 - x_2 y_1 \\
y_2 - y_3 & y_3 - y_1 & y_1 - y_2 \\
x_3 - x_2 & x_1 - x_3 & x_2 - x_1
\end{pmatrix}
$$

so that

$$
\begin{aligned}
b_1 &= \frac{1}{D}[(x_2 y_3 - x_3 y_2)u_1 + (x_3 y_1 - x_1 y_3)u_2 + (x_1 y_2 - x_2 y_1)u_3] \\
b_2 &= \frac{1}{D}[(y_2 - y_3)u_1 + (y_3 - y_1)u_2 + (y_1 - y_2)u_3] \\
b_3 &= \frac{1}{D}[(x_3 - x_2)u_1 + (x_1 - x_3)u_2 + (x_2 - x_1)u_3]
\end{aligned}
\tag{8.114}
$$

We then have

$$
\begin{aligned}
u^h(x, y) &= b_1 + b_2 x + b_3 y \\
&= \frac{1}{D}[(x_2 y_3 - x_3 y_2)u_1 + (x_3 y_1 - x_1 y_3)u_2 + (x_1 y_2 - x_2 y_1)u_3] \\
&+ \frac{1}{D}[(y_2 - y_3)u_1 + (y_3 - y_1)u_2 + (y_1 - y_2)u_3]x \\
&+ \frac{1}{D}[(x_3 - x_2)u_1 + (x_1 - x_3)u_2 + (x_2 - x_1)u_3]y \\
&= \frac{1}{D}[(x_2 y_3 - x_3 y_2) + (y_2 - y_3)x + (x_3 - x_2)y]u_1 \\
&+ \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x + (x_1 - x_3)y]u_2 \\
&+ \frac{1}{D}[(x_1 y_2 - x_2 y_1) + (y_1 - y_2)x + (x_2 - x_1)y]u_3 \\
&= \mathcal{N}_1(x, y)u_1 + \mathcal{N}_2(x, y)u_2 + \mathcal{N}_3(x, y)u_3 \tag{8.115}
\end{aligned}
$$

with the linear basis functions are given by:

$$
\begin{aligned}
\mathcal{N}_1(x, y) &= \frac{1}{D}[(x_2 y_3 - x_3 y_2) + (y_2 - y_3)x + (x_3 - x_2)y] \\
\mathcal{N}_2(x, y) &= \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x + (x_1 - x_3)y] \\
\mathcal{N}_3(x, y) &= \frac{1}{D}[(x_1 y_2 - x_2 y_1) + (y_1 - y_2)x + (x_2 - x_1)y]
\end{aligned}
$$

We can then easily verify that for example

$$
\mathcal{N}_2(x_1, y_1) = \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x_1 + (x_1 - x_3)y_1] = 0 \tag{8.116}
$$

$$
\mathcal{N}_2(x_2, y_2) = \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x_2 + (x_1 - x_3)y_2] = 1 \tag{8.117}
$$

$$
\mathcal{N}_2(x_3, y_3) = \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x_3 + (x_1 - x_3)y_3] = 0 \tag{8.118}
$$

Note that the area A of the triangle is given by:

$$
A = \frac{1}{2}D = \frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}
$$

If we now consider the reference element in the reduced coordinates space (r, s):

$(x_1, y_1) = (0, 0)$

$(x_2, y_2) = (1, 0)$

$(x_3, y_3) = (0, 1)$
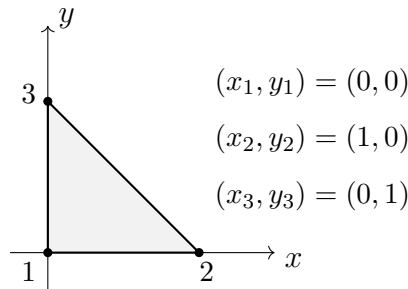
The basis polynomial is then

$$
f(r, s) = a + br + cs
$$

and the basis functions:

299

$$
\begin{aligned}
\mathcal{N}_0(r,s) &= 1 - r - s & (8.119)\\
\mathcal{N}_1(r,s) &= r & (8.120)\\
\mathcal{N}_2(r,s) &= s & (8.121)
\end{aligned}
$$

Once again we can verify that $\mathcal{N}_i(x_j, y_j) = \delta_{ij}$ and $\sum_i \mathcal{N}_i(r,s) = 1$.

8.3.11 Linear basis functions for quadrilaterals in 2D (P_1)

On the reference element $\Omega = [-1,1] \times [-1,1]$ we have three nodes placed as follows:
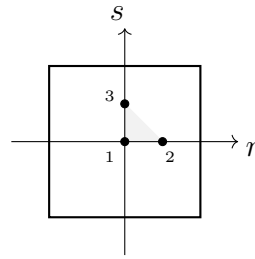
Let us assume that the function $f(r,s)$ is to be approximated on $[-1,1] \times [-1,1]$ by

$$f^h(r,s) = a + br + cs$$

Note that this is a linear function, not a bilinear one (a direct consequence of this is the fact that this function cannot be continuous from one element to another). The function f^h then must fulfill:

$$
\begin{aligned}
f^h(r_1, s_1) &= a && = f_1\\
f^h(r_2, s_2) &= a + \frac{b}{2} &&= f_2\\
f^h(r_3, s_3) &= a + \frac{c}{2} &&= f_3
\end{aligned}
$$

This leads to :

$$a = f_1 \qquad b = 2(f_2 - f_1) \qquad c = 2(f_3 - f_1)$$

Then

$$f(r,s) = f_1 + 2(f_2 - f_1)r + 2(f_3 - f_1)s$$

or,

$$f(r) = \sum_{i=1}^{3} \mathcal{N}_i(r,s) f_i$$

with

$$
\begin{aligned}
\mathcal{N}_1(r) &= 1 - 2(r+s)\\
\mathcal{N}_2(r) &= 2r\\
\mathcal{N}_3(r) &= 2s & (8.122)
\end{aligned}
$$

300

Note that we could also have placed the nodes at a different location:
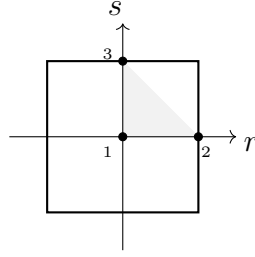
and we would then have

$$
\begin{aligned}
\mathcal{N}_1(r) &= 1 - r - s \\
\mathcal{N}_2(r) &= r \\
\mathcal{N}_3(r) &= s
\end{aligned}
\tag{8.123}
$$

as obtained in Section 8.3.10.

8.3.12 Enriched linear basis functions in triangles (P_1^+)

As we will see in Section 10.5.7 the above P_1 can be enriched with a so-called bubble function. The bubble function of the MINI element is described in Arnold *et al.* (1984) [26] as being $\lambda_1 \lambda_2 \lambda_3$ where λ_i are the so-called barycentric coordinates[9].

$$
\begin{aligned}
\lambda_1 &= \frac{(y_2 - y_3)(x - x_3) + (x_3 - x_2)(y - y_3)}{(y_2 - y_3)(x_1 - x_3) + (x_3 - x_2)(y_1 - y_3)} \\
\lambda_2 &= \frac{(y_3 - y_1)(x - x_3) + (x_1 - x_3)(y - y_3)}{(y_2 - y_3)(x_1 - x_3) + (x_3 - x_2)(y_1 - y_3)} \\
\lambda_3 &= 1 - \lambda_1 - \lambda_2
\end{aligned}
$$
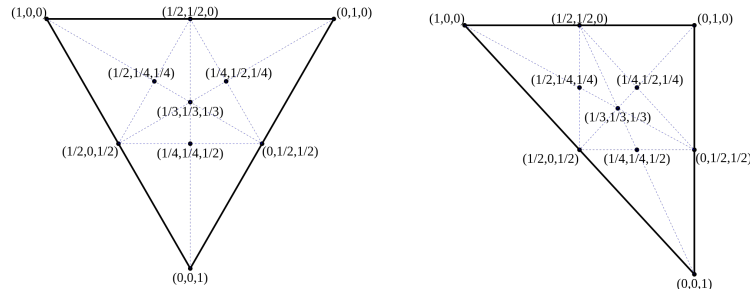
Barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$ on an equilateral triangle and on a right triangle.

---

[9] https://en.wikipedia.org/wiki/Barycentric_coordinate_system

In the reference triangle, the barycentric coordinates write

$$\lambda_1 = \frac{(s_2 - s_3)(r - r_3) + (r_3 - r_2)(s - s_3)}{(s_2 - s_3)(r_1 - r_3) + (r_3 - r_2)(s_1 - s_3)} = \frac{(-1)(r) + (-1)(s - 1)}{(-1)(0) + (-1)(-1)} = -r - s + 1$$

$$\lambda_2 = \frac{(s3 - s1)(r - r3) + (r1 - r3)(s - s3)}{(s2 - s3)(r1 - r3) + (r3 - r2)(s1 - s3)} = \frac{(1)(r) + (0)(s - 1)}{(-1)(0) + (-1)(-1)} = r$$

$$\lambda_3 = 1 - \lambda_1 - \lambda_2 = 1 - (-r - s + 1) - r = s$$

As we have seen before the bubble function is given by $\lambda_1\lambda_2\lambda_3 = (1 - r - s)rs$ and the polynomial form for the basis functions is given by:

$$f(r, s) = a + br + cs + d(1 - r - s)rs$$

Setting the location of the bubble at $r = s = 1/3$, i.e. $\lambda_1\lambda_2\lambda_3 = 1/3$, we then have

$$
\begin{aligned}
f(r_1, s_1) &= f_1 = a + br_1 + cs_1 + d(1 - r_1 - s_1)r_1 s_1 = a \\
f(r_2, s_2) &= f_2 = a + br_2 + cs_2 + d(1 - r_2 - s_2)r_2 s_2 = a + b \\
f(r_3, s_3) &= f_3 = a + br_3 + cs_3 + d(1 - r_3 - s_3)r_3 s_3 = a + c \\
f(r_4, s_4) &= f_4 = a + br_4 + cs_4 + d(1 - r_4 - s_4)r_4 s_4 = a + \frac{b}{3} + \frac{c}{3} + \frac{1}{27}
\end{aligned}
$$

where point 4 is the location of the bubble. This yields

$$a = f_1 \qquad\qquad b = f_2 - a = f_2 - f_1 \qquad\qquad c = f_3 - a = f_3 - f_1$$

and

$$d = 27\left(f_4 - a - \frac{b}{3} - \frac{c}{3}\right) = 27\left(f_4 - f_1 - \frac{f_2 - f_1}{3} - \frac{f_3 - f_1}{3}\right) = 27\left(f_4 - \frac{f_1}{3} - \frac{f_2}{3} - \frac{f_3}{3}\right)$$

Finally

$$
\begin{aligned}
f(r, s) &= a + br + cs + d(1 - r - s)rs \\
&= f_1 + (f_2 - f_1)r + (f_3 - f_1)s + 27\left(f_4 - \frac{f_1}{3} - \frac{f_2}{3} - \frac{f_3}{3}\right)(1 - r - s)rs \\
&= [1 - r - s - 9(1 - r - s)rs]f_1 + [r - 9(1 - r - s)rs]f_2 + [s - 9(1 - r - s)rs]f_3 + [27(1 - r - s)rs
\end{aligned}
$$

so that

$$f(r, s) = \sum_{i=1}^{4} \mathcal{N}_i(r, s)f_i$$

with

$$
\begin{aligned}
\mathcal{N}_1(r, s) &= 1 - r - s - 9(1 - r - s)rs \\
\mathcal{N}_2(r, s) &= r - 9(1 - r - s)rs \\
\mathcal{N}_3(r, s) &= s - 9(1 - r - s)rs \\
\mathcal{N}_4(r, s) &= 27(1 - r - s)rs
\end{aligned}
$$

It is trivial to verify that $\sum_i \mathcal{N}_i = 1$ for all values of r, s and the gradients of the basis functions are:

$$\frac{\partial \mathcal{N}_1}{\partial r}(r, s) = -1 - 9(1 - 2r - s)s \tag{8.124}$$

$$\frac{\partial \mathcal{N}_2}{\partial r}(r, s) = +1 - 9(1 - 2r - s)s \tag{8.125}$$

$$\frac{\partial \mathcal{N}_3}{\partial r}(r, s) = -9(1 - 2r - s)s \tag{8.126}$$

$$\frac{\partial \mathcal{N}_4}{\partial r}(r, s) = 27(1 - 2r - s)s \tag{8.127}$$

$$\tag{8.128}$$

$$\frac{\partial \mathcal{N}_1}{\partial s}(r, s) = -1 - 9(1 - r - 2s)r \tag{8.129}$$

$$\frac{\partial \mathcal{N}_2}{\partial s}(r, s) = -9(1 - r - 2s)r \tag{8.130}$$

$$\frac{\partial \mathcal{N}_3}{\partial s}(r, s) = +1 - 9(1 - r - 2s)r \tag{8.131}$$

$$\frac{\partial \mathcal{N}_4}{\partial s}(r, s) = 27(1 - r - 2s)r \tag{8.132}$$

We have two coordinate systems for the element: the global Cartesian coordinates (x, y) and the natural/reduced coordinates (r, s). Inside the element, the relation between the two is given by

$$
\begin{aligned}
x &= N_1 x_1 + N_2 x_2 + N_3 x_3 + N_4 x_4 = \sum_i \mathcal{N}_i(r, s) x_i \\
y &= N_1 y_1 + N_2 y_2 + N_3 y_3 + N_4 y_4 = \sum_i \mathcal{N}_i(r, s) y_i
\end{aligned}
\tag{8.133}
$$

or,

$$
\begin{aligned}
x &= [1 - r - s - 9(1 - r - s)rs]x_1 + [r - 9(1 - r - s)rs]x_2 + [s - 9(1 - r - s)rs]x_3 + [27(1 - r - s)rs]x_4 \\
&= x_1 - r(x_1 - x_2) - s(x_1 - x_3) + (1 - r - s)rs(-9x_1 - 9x_2 - 9x_3 + 27x_4) \\
&= x_1 - r(x_1 - x_2) - s(x_1 - x_3) + (1 - r - s)rs(-9x_1 - 9x_2 - 9x_3 + 27(x_1 + x_2 + x_3)/3) \\
&= x_1 - r(x_1 - x_2) - s(x_1 - x_3) \\
&= x_1 - rx_{12} - sx_{13} \\
y &= [1 - r - s - 9(1 - r - s)rs]y_1 + [r - 9(1 - r - s)rs]y_2 + [s - 9(1 - r - s)rs]y_3 + [27(1 - r - s)rs]y_4 \\
&= y_1 - r(y_1 - y_2) - s(y_1 - y_3) + (1 - r - s)rs(-9y_1 - 9y_2 - 9y_3 + 27y_4) \\
&= y_1 - r(y_1 - y_2) - s(y_1 - y_3) + (1 - r - s)rs(-9y_1 - 9y_2 - 9y_3 + 27(y_1 + y_2 + y_3)/3) \\
&= y_1 - r(y_1 - y_2) - s(y_1 - y_3) \\
&= y_1 - ry_{12} - sy_{13}
\end{aligned}
$$

8.3.13 Quadratic basis functions for triangles in 2D (P_2)

$$(r_0, s_0) = (0,0) \qquad (r_3, s_3) = (1/2, 0)$$
$$(r_1, s_1) = (1,0) \qquad (r_4, s_4) = (1/2, 1/2)$$
$$(r_2, s_2) = (0,1) \qquad (r_5, s_5) = (0, 1/2)$$
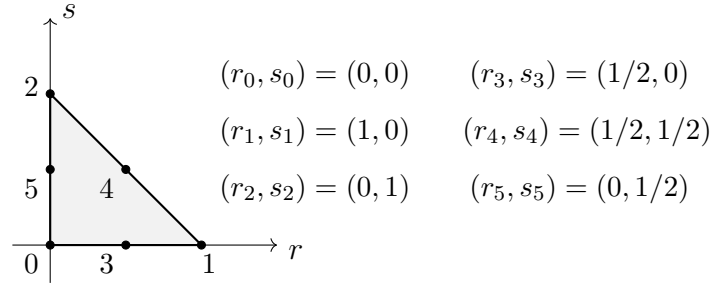
The basis polynomial is then

$$f(r,s) = c_1 + c_2 r + c_3 s + c_4 r^2 + c_5 rs + c_6 s^2$$

We have

$$
\begin{aligned}
f_0 = f(r_0, s_0) &= c_1 \\
f_1 = f(r_1, s_1) &= c_1 + c_2 + c_4 \\
f_2 = f(r_2, s_2) &= c_1 + c_3 + c_6 \\
f_3 = f(r_3, s_3) &= c_1 + c_2/2 + c_4/4 \\
f_4 = f(r_4, s_4) &= c_1 + c_2/2 + c_3/2 + c_4/4 + c_5/4 + c_6/4 \\
f_5 = f(r_5, s_5) &= c_1 + c_3/2 + c_6/4
\end{aligned}
$$

This can be cast as $\vec{f} = \boldsymbol{A} \cdot \vec{c}$ where $\boldsymbol{A}$ is a 6×6 matrix:

$$
\boldsymbol{A} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 1/2 & 0 & 1/4 & 0 & 0 \\
1 & 1/2 & 1/2 & 1/4 & 1/4 & 1/4 \\
1 & 0 & 1/2 & 0 & 0 & 1/4
\end{pmatrix}
$$

As it turns out it is rather trivial to compute the inverse of this matrix:

$$
\boldsymbol{A}^{-1} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
-3 & -1 & 0 & 4 & 0 & 0 \\
-3 & 0 & -1 & 0 & 0 & 4 \\
2 & 2 & 0 & -4 & 0 & 0 \\
4 & 0 & 0 & -4 & 4 & -4 \\
2 & 0 & 2 & 0 & 0 & -4
\end{pmatrix}
$$

Using $\vec{c} = \boldsymbol{A}^{-1} \cdot \vec{f}$ one then obtains:

$$
\begin{aligned}
c_1 &= f_0 \\
c_2 &= -3f_0 - f_1 + 4f_3 \\
c_3 &= -3f_0 - f_2 + 4f_5 \\
c_4 &= 2f_0 + 2f_1 - 4f_3 \\
c_5 &= 4f_0 - 4f_3 + 4f_4 - 4f_5 \\
c_6 &= (2f_0 + 2f_2 - 4f_5
\end{aligned}
$$

and then

$$
\begin{aligned}
f(r, s) &= f_0 + (-3f_0 - f_1 + 4f_3)r + (-3f_0 - f_2 + 4f_5)s \\
&\quad + (2f_0 + 2f_1 - 4f_3)r^2 + (4f_0 - 4f_3 + 4f_4 - 4f_5)rs + (2f_0 + 2f_2 - 4f_5)s^2 \\
&= \sum_{i=0}^{5} \mathcal{N}_i(r, s) f_i
\end{aligned}
\tag{8.134}
$$

with

$$
\begin{aligned}
\mathcal{N}_0(r, s) &= 1 - 3r - 3s + 2r^2 + 4rs + 2s^2 \\
\mathcal{N}_1(r, s) &= -r + 2r^2 \\
\mathcal{N}_2(r, s) &= -s + 2s^2 \\
\mathcal{N}_3(r, s) &= 4r - 4r^2 - 4rs \\
\mathcal{N}_4(r, s) &= 4rs \\
\mathcal{N}_5(r, s) &= 4s - 4rs - 4s^2
\end{aligned}
$$

The derivatives are as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{N}_0}{\partial r}(r, s) &= -3 + 4r + 4s \\
\frac{\partial \mathcal{N}_1}{\partial r}(r, s) &= -1 + 4r \\
\frac{\partial \mathcal{N}_2}{\partial r}(r, s) &= 0 \\
\frac{\partial \mathcal{N}_3}{\partial r}(r, s) &= 4 - 8r - 4s \\
\frac{\partial \mathcal{N}_4}{\partial r}(r, s) &= 4s \\
\frac{\partial \mathcal{N}_5}{\partial r}(r, s) &= -4s
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{N}_0}{\partial s}(r, s) &= -3 + 4r + 4s \\
\frac{\partial \mathcal{N}_1}{\partial s}(r, s) &= 0 \\
\frac{\partial \mathcal{N}_2}{\partial s}(r, s) &= -1 + 4s \\
\frac{\partial \mathcal{N}_3}{\partial s}(r, s) &= -4r \\
\frac{\partial \mathcal{N}_4}{\partial s}(r, s) &= 4r \\
\frac{\partial \mathcal{N}_5}{\partial s}(r, s) &= 4 - 4r - 8s
\end{aligned}
$$

8.3.14 Enriched quadratic basis functions in triangles (P_2^+)

This is used by the Crouzeix-Raviart element, see Section 10.5.9.

$(r_0, s_0) = (0, 0)$ $(r_3, s_3) = (1/2, 0)$ $(r_6, s_6) = (1/3, 1/3)$

$(r_1, s_1) = (1, 0)$ $(r_4, s_4) = (1/2, 1/2)$

$(r_2, s_2) = (0, 1)$ $(r_5, s_5) = (0, 1/2)$
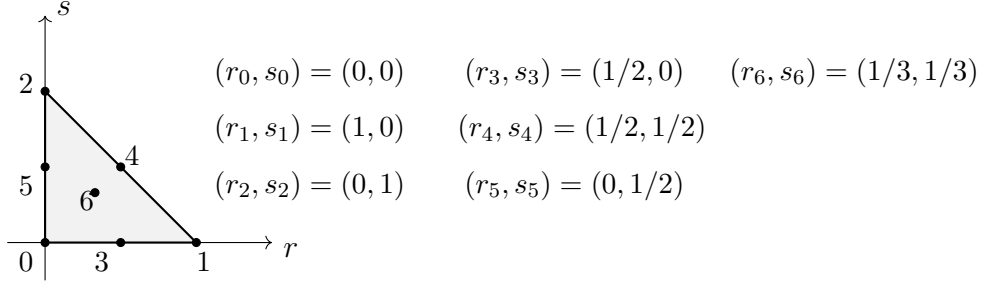
The basis functions are given by:

find reference

$$
\begin{aligned}
\mathcal{N}_0(r, s) &= (1 - r - s)(1 - 2r - 2s + 3rs) & (8.135) \\
\mathcal{N}_1(r, s) &= r(2r - 1 + 3s - 3rs - 3s^2) & (8.136) \\
\mathcal{N}_2(r, s) &= s(2s - 1 + 3r - 3r^2 - 3rs) & (8.137) \\
\mathcal{N}_3(r, s) &= 4(1 - r - s)r(1 - 3s) & (8.138) \\
\mathcal{N}_4(r, s) &= 4rs[-2 + 3r + 3s] & (8.139) \\
\mathcal{N}_5(r, s) &= 4(1 - r - s)s(1 - 3r) & (8.140) \\
\mathcal{N}_6(r, s) &= 27(1 - r - s)rs & (8.141)
\end{aligned}
$$

It is then easy to verify that for all basis functions we have $\mathcal{N}_i(r_j, s_j) = \delta_{ij}$ where j denotes one of the seven nodes. The derivatives are as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{N}_0}{\partial r}(r, s) &= r(4 - 6s) - 3s^2 + 7s - 3 & (8.142) \\
\frac{\partial \mathcal{N}_1}{\partial r}(r, s) &= r(4 - 6s) - 3s^2 + 3s - 1 & (8.143) \\
\frac{\partial \mathcal{N}_2}{\partial r}(r, s) &= -3s(2r + s - 1) & (8.144) \\
\frac{\partial \mathcal{N}_3}{\partial r}(r, s) &= 4(3s - 1)(2r + s - 1) & (8.145) \\
\frac{\partial \mathcal{N}_4}{\partial r}(r, s) &= 4s(6r + 3s - 2) & (8.146) \\
\frac{\partial \mathcal{N}_5}{\partial r}(r, s) &= 4s(6r + 3s - 4) & (8.147) \\
\frac{\partial \mathcal{N}_6}{\partial r}(r, s) &= -27s(2r + s - 1) & (8.148)
\end{aligned}
$$

$$\frac{\partial \mathcal{N}_0}{\partial s}(r,s) \;=\; -3r^2 + r(7-6s) + 4s - 3 \tag{8.149}$$

$$\frac{\partial \mathcal{N}_1}{\partial s}(r,s) \;=\; -3r(r+2s-1) \tag{8.150}$$

$$\frac{\partial \mathcal{N}_2}{\partial s}(r,s) \;=\; -3r^2 + r(3-6s) + 4s - 1 \tag{8.151}$$

$$\frac{\partial \mathcal{N}_3}{\partial s}(r,s) \;=\; 4r(3r+6s-4) \tag{8.152}$$

$$\frac{\partial \mathcal{N}_4}{\partial s}(r,s) \;=\; 4r(3r+6s-2) \tag{8.153}$$

$$\frac{\partial \mathcal{N}_5}{\partial s}(r,s) \;=\; 4(3r-1)(r+2s-1) \tag{8.154}$$

$$\frac{\partial \mathcal{N}_6}{\partial s}(r,s) \;=\; -27r(r+2s-1) \tag{8.155}$$

Note that the basis functions can also be expressed as a function of the barycentric coordinates, as in the MILAMIN code [318] or in Cuvelier *et al.* (1986) [317][10]

$$N_0(\lambda_1, \lambda_2, \lambda_3) \;=\; \eta_1(2\eta_1 - 1) + 3\eta_1\eta_2\eta_3 \tag{8.156}$$

$$N_1(\lambda_1, \lambda_2, \lambda_3) \;=\; \eta_2(2\eta_2 - 1) + 3\eta_1\eta_2\eta_3 \tag{8.157}$$

$$N_2(\lambda_1, \lambda_2, \lambda_3) \;=\; \eta_3(2\eta_3 - 1) + 3\eta_1\eta_2\eta_3 \tag{8.158}$$

$$N_3(\lambda_1, \lambda_2, \lambda_3) \;=\; 4\eta_2\eta_3 - 12\eta_1\eta_2\eta_3 \tag{8.159}$$

$$N_4(\lambda_1, \lambda_2, \lambda_3) \;=\; 4\eta_1\eta_3 - 12\eta_1\eta_2\eta_3 \tag{8.160}$$

$$N_5(\lambda_1, \lambda_2, \lambda_3) \;=\; 4\eta_1\eta_2 - 12\eta_1\eta_2\eta_3 \tag{8.161}$$

$$N_6(\lambda_1, \lambda_2, \lambda_3) \;=\; 27\eta_1\eta_2\eta_3 \tag{8.162}$$

VERIFY that when $\eta_1 = 1 - r - s$, $\eta_2 = r$ and $\eta_3 = s$ we find the above r,s basis functions

----

[10]Note that the numbering of the nodes in the book is different with respect to the one above.

8.3.15 Cubic basis functions for triangles (P_3)

This finite element space consists of continuous piecewise cubic functions. The functionals in a mesh cell are defined to be the values in the vertices ($d + 1 = 2 + 1 = 3$ values), two values on each edge (dividing the edge in three parts of equal length) (3*2 values), and the value in the barycenter of the cell. It is implemented in STONE 120. See also python code in `images/basis_P3` which I wrote to test these basis functions.

$$(r_0, s_0) = (0, 0) \qquad (r_5, s_5) = (1/3, 1/3)$$
$$(r_1, s_1) = (1/3, 0) \qquad (r_6, s_6) = (2/3, 1/3)$$
$$(r_2, s_2) = (2/3, 0) \qquad (r_7, s_7) = (0, 2/3)$$
$$(r_3, s_3) = (1, 0) \qquad (r_8, s_8) = (1/3, 2/3)$$
$$(r_4, s_4) = (0, 1/3) \qquad (r_9, s_9) = (0, 1)$$
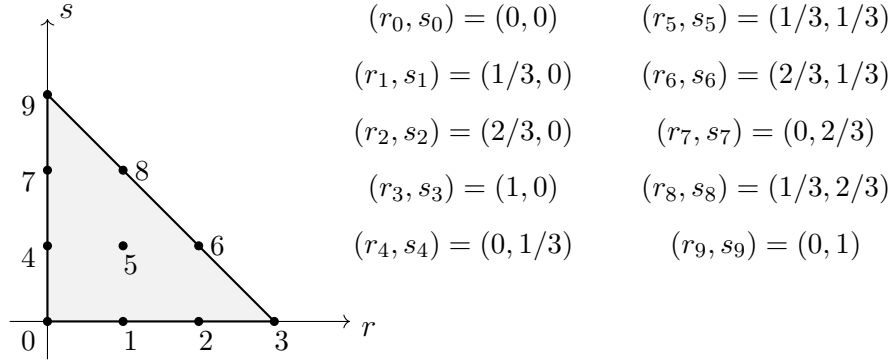
The basis polynomial is in this case

$$f(r, s) = c_0 + c_1 r + c_2 s + c_3 r^2 + c_4 rs + c_5 s^2 + c_6 r^3 + c_7 r^2 s + c_8 rs^2 + c_9 s^3$$

with the requirement that it should be such that $f(r_i, s_i) = f_i$ at the 10 nodes:

$$f_0 = f(r_0, s_0) = c_0$$
$$f_1 = f(r_1, s_1) = c_0 + c_1 + \frac{1}{9}c_3 + \frac{1}{27}c_6$$
$$f_2 = f(r_2, s_2) = c_0 + \frac{2}{3}c_1 + \frac{4}{9}c_3 + \frac{8}{27}c_6$$
$$f_3 = f(r_3, s_3) = c_0 + c_1 + c_3 + c_6$$
$$f_4 = f(r_4, s_4) = c_0 + \frac{1}{3}c_2 + \frac{1}{9}c_5 + \frac{1}{27}c_9$$
$$f_5 = f(r_5, s_5) = c_0 + \frac{1}{3}c_1 + \frac{1}{3}c_2 + \frac{1}{9}c_3 + \frac{1}{9}c_4 + \frac{1}{9}c_5 + \frac{1}{27}c_6 r^3 + \frac{1}{27}c_7 + \frac{1}{27}c_8 + \frac{1}{27}c_9$$
$$f_6 = f(r_6, s_6) = c_0 + \frac{2}{3}c_1 + \frac{1}{3}c_2 + \frac{4}{9}c_3 + \frac{2}{9}c_4 + \frac{1}{9}c_5 + \frac{8}{27}c_6 + \frac{4}{27}c_7 + \frac{2}{27}c_8 + \frac{1}{9}c_9$$
$$f_7 = f(r_7, s_7) = c_0 + \frac{2}{3}c_2 + \frac{4}{9}c_5 + \frac{8}{27}c_9$$
$$f_8 = f(r_8, s_8) = c_0 + \frac{1}{3}c_1 + \frac{2}{3}c_2 + \frac{1}{9}c_3 + \frac{2}{9}c_4 + \frac{4}{9}c_5 + \frac{1}{27}c_6 + \frac{2}{27}c_7 + \frac{4}{27}c_8 + \frac{8}{27}c_9 s^3$$
$$f_9 = f(r_9, s_9) = c_0 + c_2 + c_5 + c_9$$

or,

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & \frac{1}{3} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{27} & 0 & 0 & 0 \\
1 & \frac{2}{3} & 0 & \frac{4}{9} & 0 & 0 & \frac{8}{27} & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{9} & 0 & 0 & 0 & \frac{1}{27} \\
1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{27} & \frac{1}{27} & \frac{1}{27} & \frac{1}{27} \\
1 & \frac{2}{3} & \frac{1}{3} & \frac{4}{9} & \frac{2}{9} & \frac{1}{9} & \frac{8}{27} & \frac{4}{27} & \frac{2}{27} & \frac{1}{27} \\
1 & 0 & \frac{2}{3} & 0 & 0 & \frac{4}{9} & 0 & 0 & 0 & \frac{8}{27} \\
1 & \frac{1}{3} & \frac{2}{3} & \frac{1}{9} & \frac{2}{9} & \frac{4}{9} & \frac{1}{27} & \frac{2}{27} & \frac{4}{27} & \frac{8}{27} \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}
\cdot
\begin{pmatrix}
c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9
\end{pmatrix}
=
\begin{pmatrix}
f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9
\end{pmatrix}
$$

or,

$$
\frac{1}{27}
\begin{pmatrix}
27 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
27 & 9 & 0 & 3 & 0 & 0 & 1 & 0 & 0 & 0 \\
27 & 18 & 0 & 12 & 0 & 0 & 8 & 0 & 0 & 0 \\
27 & 27 & 0 & 27 & 0 & 0 & 27 & 0 & 0 & 0 \\
27 & 0 & 9 & 0 & 0 & 3 & 0 & 0 & 0 & 1 \\
27 & 9 & 9 & 3 & 3 & 3 & 1 & 1 & 1 & 1 \\
27 & 18 & 9 & 12 & 6 & 3 & 8 & 4 & 2 & 1 \\
27 & 0 & 18 & 0 & 0 & 12 & 0 & 0 & 0 & 8 \\
27 & 9 & 18 & 3 & 6 & 12 & 1 & 2 & 4 & 8 \\
27 & 0 & 27 & 0 & 0 & 27 & 0 & 0 & 0 & 27
\end{pmatrix}
\cdot
\begin{pmatrix}
c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9
\end{pmatrix}
=
\begin{pmatrix}
f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9
\end{pmatrix}
$$

The inverse of the matrix is

$$
\frac{1}{2}
\begin{pmatrix}
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-11 & 18 & -9 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
-11 & 0 & 0 & 0 & 18 & 0 & 0 & -9 & 0 & 2 \\
18 & -45 & 36 & -9 & 0 & 0 & 0 & 0 & 0 & 0 \\
36 & -45 & 9 & 0 & -45 & 54 & -9 & 9 & -9 & 0 \\
18 & 0 & 0 & 0 & -45 & 0 & 0 & 36 & 0 & -9 \\
-9 & 27 & -27 & 9 & 0 & 0 & 0 & 0 & 0 & 0 \\
-27 & 54 & -27 & 0 & 27 & -54 & 27 & 0 & 0 & 0 \\
-27 & 27 & 0 & 0 & 54 & -54 & 0 & -27 & 27 & 0 \\
-9 & 0 & 0 & 0 & 27 & 0 & 0 & -27 & 0 & 9
\end{pmatrix}
$$

so that the solution of the system $\boldsymbol{A} \cdot \vec{c} = \vec{f}$ is $\vec{c} = \boldsymbol{A}^{-1} \cdot \vec{f}$, or:

$$
\begin{aligned}
c_0 &= 1 \\
c_1 &= \frac{1}{2}(-11f_0 + 18f_1 - 9f_2 + 2f_3) \\
c_2 &= \frac{1}{2}(-11f_0 + 18f_4 - 9f_7 + 2f_9) \\
c_3 &= \text{etc ...}
\end{aligned}
\tag{8.163}
$$

which we insert in

$$
f(r, s) = c_0 + c_1 r + c_2 s + c_3 r^2 + c_4 r s + c_5 s^2 + c_6 r^3 + c_7 r^2 s + c_8 r s^2 + c_9 s^3
$$

309

and we then obtain

$$
\begin{aligned}
f(r,s) \;=\; & \frac{1}{2}\left(2 - 11r - 11s + 18r^2 + 36rs + 18s^2 - 9r^3 - 27r^2 s - 27rs^2 - 9s^3\right) f_0 \\
+ & \frac{1}{2}\left(18r - 45r^2 - 45rs + 27r^3 + 54r^2 s + 27rs^2\right) f_1 \\
+ & \frac{1}{2}\left(-9r + 36r^2 + 9rs - 27r^3 - 27r^2 s\right) f_2 \\
+ & \frac{1}{2}\left(2r - 9r^2 + 9r^3\right) f_3 \\
+ & \frac{1}{2}\left(18s - 45rs - 45s^2 + 27r^2 s + 54rs^2 + 27s^3\right) f_4 \\
+ & \frac{1}{2}\left(54rs - 54r^2 s - 54rs^2\right) f_5 \\
+ & \frac{1}{2}\left(-9rs + 27r^2 s\right) f_6 \\
+ & \frac{1}{2}\left(-9s + 9rs + 36s^2 - 27rs^2 - 27s^3\right) f_7 \\
+ & \frac{1}{2}\left(-9rs + 27rs^2\right) f_8 \\
+ & \frac{1}{2}\left(2s - 9s^2 + 9s^3\right) f_9 \\
\;=\; & \sum_{i=0}^{9} \mathcal{N}_i(r,s) f_i
\end{aligned}
$$

with

$$
\begin{aligned}
\mathcal{N}_0(r,s) &= \frac{1}{2}\left(2 - 11r - 11s + 18r^2 + 36rs + 18s^2 - 9r^3 - 27r^2 s - 27rs^2 - 9s^3\right) \\
\mathcal{N}_1(r,s) &= \frac{1}{2}\left(18r - 45r^2 - 45rs + 27r^3 + 54r^2 s + 27rs^2\right) \\
\mathcal{N}_2(r,s) &= \frac{1}{2}\left(-9r + 36r^2 + 9rs - 27r^3 - 27r^2 s\right) \\
\mathcal{N}_3(r,s) &= \frac{1}{2}\left(2r - 9r^2 + 9r^3\right) \\
\mathcal{N}_4(r,s) &= \frac{1}{2}\left(18s - 45rs - 45s^2 + 27r^2 s + 54rs^2 + 27s^3\right) \\
\mathcal{N}_5(r,s) &= \frac{1}{2}\left(54rs - 54r^2 s - 54rs^2\right) \\
\mathcal{N}_6(r,s) &= \frac{1}{2}\left(-9rs + 27r^2 s\right) \\
\mathcal{N}_7(r,s) &= \frac{1}{2}\left(-9s + 9rs + 36s^2 - 27rs^2 - 27s^3\right) \\
\mathcal{N}_8(r,s) &= \frac{1}{2}\left(-9rs + 27rs^2\right) \\
\mathcal{N}_9(r,s) &= \frac{1}{2}\left(2s - 9s^2 + 9s^3\right) \tag{8.164}
\end{aligned}
$$

and

$$\frac{\partial \mathcal{N}_0}{\partial r}(r,s) = \frac{1}{2}(-11 + 36r + 36s - 27r^2 - 54rs - 27s^2)$$

$$\frac{\partial \mathcal{N}_1}{\partial r}(r,s) = \frac{1}{2}(18 - 90r - 45s + 81r^2 + 108rs + 27s^2)$$

$$\frac{\partial \mathcal{N}_2}{\partial r}(r,s) = \frac{1}{2}(-9 + 72r + 9s - 81r^2 - 54rs)$$

$$\frac{\partial \mathcal{N}_3}{\partial r}(r,s) = \frac{1}{2}(2 - 18r + 27r^2)$$

$$\frac{\partial \mathcal{N}_4}{\partial r}(r,s) = \frac{1}{2}(-45s + 54rs + 54s^2)$$

$$\frac{\partial \mathcal{N}_5}{\partial r}(r,s) = \frac{1}{2}(54s - 108rs - 54s^2)$$

$$\frac{\partial \mathcal{N}_6}{\partial r}(r,s) = \frac{1}{2}(-9s + 54rs)$$

$$\frac{\partial \mathcal{N}_7}{\partial r}(r,s) = \frac{1}{2}(9s - 27s^2)$$

$$\frac{\partial \mathcal{N}_8}{\partial r}(r,s) = \frac{1}{2}(-9s + 27s^2)$$

$$\frac{\partial \mathcal{N}_9}{\partial r}(r,s) = 0$$

$$\frac{\partial \mathcal{N}_0}{\partial s}(r,s) = \frac{1}{2}(-11 + 36r + 36s - 27r^2 - 54rs - 27s^2)$$

$$\frac{\partial \mathcal{N}_1}{\partial s}(r,s) = \frac{1}{2}(-45r + 54r^2 + 54rs)$$

$$\frac{\partial \mathcal{N}_2}{\partial s}(r,s) = \frac{1}{2}(9r - 27r^2)$$

$$\frac{\partial \mathcal{N}_3}{\partial s}(r,s) = 0$$

$$\frac{\partial \mathcal{N}_4}{\partial s}(r,s) = \frac{1}{2}(18 - 45r - 90s + 27r^2 + 108rs + 81s^2)$$

$$\frac{\partial \mathcal{N}_5}{\partial s}(r,s) = \frac{1}{2}(54r - 54r^2 - 108rs)$$

$$\frac{\partial \mathcal{N}_6}{\partial s}(r,s) = \frac{1}{2}(-9r + 27r^2)$$

$$\frac{\partial \mathcal{N}_7}{\partial s}(r,s) = \frac{1}{2}(-9 + 9r + 72s - 54rs - 81s^2)$$

$$\frac{\partial \mathcal{N}_8}{\partial s}(r,s) = \frac{1}{2}(-9r + 54rs)$$

$$\frac{\partial \mathcal{N}_9}{\partial s}(r,s) = \frac{1}{2}(2 - 18s + 27s^2)$$

In barycentric coordinates the basis functions are:

- on the nodes: $\frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2)$

- on the edges: $\frac{9}{2}\lambda_i\lambda_j(3\lambda_i - 1)$

- in the middle: $27\lambda_i\lambda_j\lambda_k$

Probably found in John's book - need to check that!

8.3.16 Quartic basis functions for triangles (P_4)

It is implemented in STONE 120. See also python code in `images/basis_P4` which I wrote to test these basis functions.
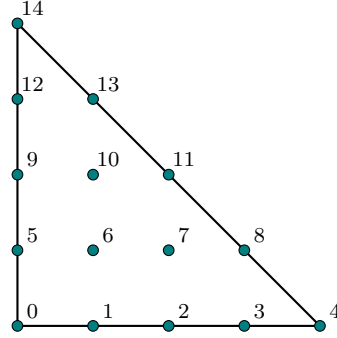
The support nodes coordinates are as follows:

$i \to$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_i	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0
s_i	0	0	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{3}{4}$	1

Inside the element a field f is represented by a 4-th order polynomial:

$$
\begin{aligned}
f^h(r,s) \; = \; & c_0 + c_1 r + c_2 s \\
& + c_3 r^2 + c_4 rs + c_5 s^2 \\
& + c_6 r^3 + c_7 r^2 s + c_8 rs^2 + c_9 s^3 \\
& + c_{10} r^4 + c_{11} r^3 s + c_{12} r^2 s^2 + c_{13} rs^3 + c_{14} s^4
\end{aligned}
\tag{8.165}
$$

At each node the function takes a value f_i, $i \in [0, 14]$ so that we have:

$$
\left(
\begin{array}{ccccccccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & \frac{1}{4} & 0 & \frac{1}{16} & 0 & 0 & \frac{1}{64} & 0 & 0 & 0 & \frac{1}{256} & 0 & 0 & 0 & 0 \\
1 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \\
1 & \frac{3}{4} & 0 & \frac{9}{16} & 0 & 0 & \frac{27}{64} & 0 & 0 & 0 & \frac{81}{256} & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\[4pt]
1 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & \frac{1}{64} & 0 & 0 & 0 & 0 & \frac{1}{256} \\
1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{64} & \frac{1}{64} & \frac{1}{64} & \frac{1}{64} & \frac{1}{256} & \frac{1}{256} & \frac{1}{256} & \frac{1}{256} & \frac{1}{256} \\
1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & \frac{1}{32} & \frac{1}{64} & \frac{1}{16} & \frac{1}{32} & \frac{1}{64} & \frac{1}{128} & \frac{1}{256} \\
1 & \frac{3}{4} & \frac{1}{4} & \frac{9}{16} & \frac{3}{16} & \frac{1}{16} & \frac{27}{64} & \frac{9}{64} & \frac{3}{64} & \frac{1}{64} & \frac{81}{256} & \frac{27}{256} & \frac{9}{256} & \frac{3}{256} & \frac{1}{256} \\[4pt]
1 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{1}{16} \\
1 & \frac{1}{4} & \frac{1}{2} & \frac{1}{16} & \frac{1}{8} & \frac{1}{4} & \frac{1}{64} & \frac{1}{32} & \frac{1}{16} & \frac{1}{8} & \frac{1}{256} & \frac{1}{128} & \frac{1}{64} & \frac{1}{32} & \frac{1}{16} \\
1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \\[4pt]
1 & 0 & \frac{3}{4} & 0 & 0 & \frac{9}{16} & 0 & 0 & 0 & \frac{27}{64} & 0 & 0 & 0 & 0 & \frac{81}{256} \\
1 & \frac{1}{4} & \frac{3}{4} & \frac{1}{16} & \frac{3}{16} & \frac{9}{16} & \frac{1}{64} & \frac{3}{64} & \frac{9}{64} & \frac{27}{64} & \frac{1}{256} & \frac{3}{256} & \frac{9}{256} & \frac{27}{256} & \frac{81}{256} \\[4pt]
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{array}
\right)
\cdot
\left(
\begin{array}{c}
c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9 \\ c_{10} \\ c_{11} \\ c_{12} \\ c_{13} \\ c_{14}
\end{array}
\right)
=
\left(
\begin{array}{c}
f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \\ f_{10} \\ f_{11} \\ f_{12} \\ f_{13} \\ f_{14}
\end{array}
\right)
$$

or,

$$\frac{1}{256}\begin{pmatrix}
256 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
256 & 64 & 0 & 16 & 0 & 0 & 4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
256 & 128 & 0 & 64 & 0 & 0 & 32 & 0 & 0 & 0 & 16 & 0 & 0 & 0 & 0 \\
256 & 192 & 0 & 144 & 0 & 0 & 108 & 0 & 0 & 0 & 81 & 0 & 0 & 0 & 0 \\
256 & 256 & 0 & 256 & 0 & 0 & 256 & 0 & 0 & 0 & 256 & 0 & 0 & 0 & 0 \\
256 & 0 & 64 & 0 & 0 & 16 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 1 \\
256 & 64 & 64 & 16 & 16 & 16 & 4 & 4 & 4 & 4 & 1 & 1 & 1 & 1 & 1 \\
256 & 128 & 64 & 64 & 32 & 16 & 32 & 16 & 8 & 4 & 16 & 8 & 4 & 2 & 1 \\
256 & 192 & 64 & 144 & 48 & 16 & 108 & 36 & 12 & 4 & 81 & 27 & 9 & 3 & 1 \\
256 & 0 & 128 & 0 & 0 & 64 & 0 & 0 & 0 & 32 & 0 & 0 & 0 & 0 & 16 \\
256 & 64 & 128 & 16 & 32 & 64 & 4 & 8 & 16 & 32 & 1 & 2 & 4 & 8 & 16 \\
256 & 128 & 128 & 64 & 64 & 64 & 32 & 32 & 32 & 32 & 16 & 16 & 16 & 16 & 16 \\
256 & 0 & 192 & 0 & 0 & 144 & 0 & 0 & 0 & 108 & 0 & 0 & 0 & 0 & 81 \\
256 & 64 & 192 & 16 & 48 & 144 & 4 & 12 & 36 & 108 & 1 & 3 & 9 & 27 & 81 \\
256 & 0 & 256 & 0 & 0 & 256 & 0 & 0 & 0 & 256 & 0 & 0 & 0 & 0 & 256
\end{pmatrix}\cdot\begin{pmatrix}
c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9 \\ c_{10} \\ c_{11} \\ c_{12} \\ c_{13} \\ c_{14}
\end{pmatrix}=\begin{pmatrix}
f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f \\ f
\end{pmatrix}$$

The inverse of the matrix is:

$$\frac{1}{3}\begin{pmatrix}
3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-25 & 48 & -36 & 16 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-25 & 0 & 0 & 0 & 0 & 48 & 0 & 0 & 0 & -36 & 0 & 0 & 16 & 0 & -3 \\
70 & -208 & 228 & -112 & 22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
140 & -208 & 84 & -16 & 0 & -208 & 288 & -96 & 16 & 84 & -96 & 12 & -16 & 16 & 0 \\
70 & 0 & 0 & 0 & 0 & -208 & 0 & 0 & 0 & 228 & 0 & 0 & -112 & 0 & 22 \\
-80 & 288 & -384 & 224 & -48 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-240 & 576 & -432 & 96 & 0 & 288 & -672 & 480 & -96 & -48 & 96 & -48 & 0 & 0 & 0 \\
-240 & 288 & -48 & 0 & 0 & 576 & -672 & 96 & 0 & -432 & 480 & -48 & 96 & -96 & 0 \\
-80 & 0 & 0 & 0 & 0 & 288 & 0 & 0 & 0 & -384 & 0 & 0 & 224 & 0 & -48 \\
32 & -128 & 192 & -128 & 32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
128 & -384 & 384 & -128 & 0 & -128 & 384 & -384 & 128 & 0 & 0 & 0 & 0 & 0 & 0 \\
192 & -384 & 192 & 0 & 0 & -384 & 768 & -384 & 0 & 192 & -384 & 192 & 0 & 0 & 0 \\
128 & -128 & 0 & 0 & 0 & -384 & 384 & 0 & 0 & 384 & -384 & 0 & -128 & 128 & 0 \\
32 & 0 & 0 & 0 & 0 & -128 & 0 & 0 & 0 & 192 & 0 & 0 & -128 & 0 & 32
\end{pmatrix}$$

313

so that we obtain

$$
\begin{aligned}
f(r,s) \;=\; & \tfrac{1}{3}(3 - 25r - 25s + 70r^2 + 140rs + 70s^2 - 80r^3 - 240r^2s - 240rs^2 - 80s^3 + 32r^4 + 128r^3s + 192r^2s^2 + 128rs^3 + 32s^4)f_0 \\[4pt]
& + \tfrac{1}{3}(48r - 208r^2 - 208rs + 288r^3 + 576r^2s + 288rs^2 - 128r^4 - 384r^3s - 384r^2s^2 - 128rs^3)f_1 \\[4pt]
& + \tfrac{1}{3}(-36r + 228r^2 + 84rs - 384r^3 - 432r^2s - 48rs^2 + 192r^4 + 384r^3s + 192r^2s^2)f_2 \\[4pt]
& + \tfrac{1}{3}(16r - 112r^2 - 16rs + 224r^3 + 96r^2s - 128r^4 - 128r^3s)f_3 \\[4pt]
& + \tfrac{1}{3}(-3r + 22r^2 - 48r^3 + 32r^4)f_4 \\[4pt]
& + \tfrac{1}{3}(48s - 208rs - 208s^2 + 288r^2s + 576rs^2 + 288s^3 - 128r^3s - 384r^2s^2 - 384rs^3 - 128s^4)f_5 \\[4pt]
& + \tfrac{1}{3}(288rs - 672r^2s - 672rs^2 + 384r^3s + 768r^2s^2 + 384rs^3)f_6 \\[4pt]
& + \tfrac{1}{3}(-96rs + 480r^2s + 96rs^2 - 384r^3s - 384r^2s^2)f_7 \\[4pt]
& + \tfrac{1}{3}(16rs - 96r^2s + 128r^3s)f_8 \\[4pt]
& + \tfrac{1}{3}(-36s + 84rs + 228s^2 - 48r^2s - 432rs^2 - 384s^3 + 192r^2s^2 + 384rs^3 + 192s^4)f_9 \\[4pt]
& + \tfrac{1}{3}(-96rs + 96r^2s + 480rs^2 - 384r^2s^2 - 384rs^3)f_{10} \\[4pt]
& + \tfrac{1}{3}(12rs - 48r^2s - 48rs^2 + 192r^2s^2)f_{11} \\[4pt]
& + \tfrac{1}{3}(16s - 16rs - 112s^2 + 96rs^2 + 224s^3 - 128rs^3 - 128s^4)f_{12} \\[4pt]
& + \tfrac{1}{3}(16rs - 96rs^2 + 128rs^3)f_{13} \\[4pt]
& + \tfrac{1}{3}(-3s + 22s^2 - 48s^3 + 32s^4)f_{14} \\[4pt]
\;=\; & \sum_{i=0}^{14} \mathcal{N}_i(r,s)f_i
\end{aligned}
\tag{8.166}
$$

and then

$$\frac{\partial \mathcal{N}_0}{\partial r}(r,s) = \frac{1}{3}(-25 + 140r + 140s - 240r^2 - 480rs - 240s^2 + 128r^3 + 384r^2s + 384rs^2 + 128s^3)$$

$$\frac{\partial \mathcal{N}_1}{\partial r}(r,s) = \frac{1}{3}(48 - 416r - 208s + 864r^2 + 1152rs + 288s^2 - 512r^3 - 1152r^2s - 768rs^2 - 128s^3)$$

$$\frac{\partial \mathcal{N}_2}{\partial r}(r,s) = \frac{1}{3}(-36 + 456r + 84s - 1152r^2 - 864rs - 48s^2 + 768r^3 + 1152r^2s + 384rs^2)$$

$$\frac{\partial \mathcal{N}_3}{\partial r}(r,s) = \frac{1}{3}(16 - 224r - 16s + 672r^2 + 192rs - 512r^3 - 384r^2s)$$

$$\frac{\partial \mathcal{N}_4}{\partial r}(r,s) = \frac{1}{3}(-3 + 44r - 144r^2 + 128r^3)$$

$$\frac{\partial \mathcal{N}_5}{\partial r}(r,s) = \frac{1}{3}(-208s + 576rs + 576s^2 - 384r^2s - 768rs^2 - 384s^3)$$

$$\frac{\partial \mathcal{N}_6}{\partial r}(r,s) = \frac{1}{3}(288s - 1344rs - 672s^2 + 1152r^2s + 1536rs^2 + 384s^3)$$

$$\frac{\partial \mathcal{N}_7}{\partial r}(r,s) = \frac{1}{3}(-96s + 960rs + 96s^2 - 1152r^2s - 768rs^2)$$

$$\frac{\partial \mathcal{N}_8}{\partial r}(r,s) = \frac{1}{3}(16s - 192rs + 384r^2s)$$

$$\frac{\partial \mathcal{N}_9}{\partial r}(r,s) = \frac{1}{3}(84s - 96rs - 432s^2 + 384rs^2 + 384s^3)$$

$$\frac{\partial \mathcal{N}_{10}}{\partial r}(r,s) = \frac{1}{3}(-96s + 192rs + 480s^2 - 768rs^2 - 384s^3)$$

$$\frac{\partial \mathcal{N}_{11}}{\partial r}(r,s) = \frac{1}{3}(12s - 96rs - 48s^2 + 384rs^2)$$

$$\frac{\partial \mathcal{N}_{12}}{\partial r}(r,s) = \frac{1}{3}(-16s + 96s^2 - 128s^3)$$

$$\frac{\partial \mathcal{N}_{13}}{\partial r}(r,s) = \frac{1}{3}(16s - 96s^2 + 128s^3)$$

$$\frac{\partial \mathcal{N}_{14}}{\partial r}(r,s) = 0$$

$$\frac{\partial \mathcal{N}_0}{\partial s}(r,s) = \frac{1}{3}(-25 + 140r + 140s - 240r^2 - 480rs - 240s^2 + 128r^3 + 384r^2s + 384rs^2 + 128s^3)$$

$$\frac{\partial \mathcal{N}_1}{\partial s}(r,s) = \frac{1}{3}(-208r + 576r^2 + 576rs - 384r^3 - 768r^2s - 384rs^2)$$

$$\frac{\partial \mathcal{N}_2}{\partial s}(r,s) = \frac{1}{3}(84r - 432r^2 - 96rs + 384r^3 + 384r^2s)$$

$$\frac{\partial \mathcal{N}_3}{\partial s}(r,s) = \frac{1}{3}(-16r + 96r^2 - 128r^3)$$

$$\frac{\partial \mathcal{N}_4}{\partial s}(r,s) = 0$$

$$\frac{\partial \mathcal{N}_5}{\partial s}(r,s) = \frac{1}{3}(48 - 208r - 416s + 288r^2 + 1152rs + 864s^2 - 128r^3 - 768r^2s - 1152rs^2 - 512s^3)$$

$$\frac{\partial \mathcal{N}_6}{\partial s}(r,s) = \frac{1}{3}(288r - 672r^2 - 1344rs + 384r^3 + 1536r^2s + 1152rs^2)$$

$$\frac{\partial \mathcal{N}_7}{\partial s}(r,s) = \frac{1}{3}(-96r + 480r^2 + 192rs - 384r^3 - 768r^2s)$$

$$\frac{\partial \mathcal{N}_8}{\partial s}(r,s) = \frac{1}{3}(16r - 96r^2 + 128r^3)$$

$$\frac{\partial \mathcal{N}_9}{\partial s}(r,s) = \frac{1}{3}(-36 + 84r + 456s - 48r^2 - 864rs - 1152s^2 + 384r^2s + 1152rs^2 + 768s^3)$$

$$\frac{\partial \mathcal{N}_{10}}{\partial s}(r,s) = \frac{1}{3}(-96r + 96r^2 + 960rs - 768r^2s - 1152rs^2)$$

$$\frac{\partial \mathcal{N}_{11}}{\partial s}(r,s) = \frac{1}{3}(12r - 48r^2 - 96rs + 384r^2s)$$

$$\frac{\partial \mathcal{N}_{12}}{\partial s}(r,s) = \frac{1}{3}(16 - 16r - 224s + 192rs + 672s^2 - 384rs^2 - 512s^3)$$

$$\frac{\partial \mathcal{N}_{13}}{\partial s}(r,s) = \frac{1}{3}(16r - 192rs + 384rs^2)$$

$$\frac{\partial \mathcal{N}_{14}}{\partial s}(r,s) = \frac{1}{3}(-3 + 44s - 144s^2 + 128s^3)$$

8.3.17 The P_1^{NC} space

```
P1         P1NC
2          .               (r0,s0)=(1/2,0)
| \        | \
|  \       2  1            (r1,s1)=(1/2,1/2)
|   \      |   \
|    \     |    \
0----1    .--0--.          (r2,s2)=(0,1/2)
```

The basis functions are $1 - 2\lambda_i$, where λ_i are the barycentric coordinates, so we arrive at

$$\begin{aligned}
\mathcal{N}_0(r,s) &= 1 - 2\lambda_3 = 1 - 2s \\
\mathcal{N}_1(r,s) &= 1 - 2\lambda_1 = 1 - 2(1 - r - s) = -1 + 2r + 2s \\
\mathcal{N}_2(r,s) &= 1 - 2\lambda_2 = 1 - 2r
\end{aligned}$$

with of course

$$\mathcal{N}_0(r,s) + \mathcal{N}_1(r,s) + \mathcal{N}_2(r,s) = (1 - 2s) + (-1 + 2r + 2s) + (1 - 2r) = 1$$

316

We have

$$
\begin{array}{rcl}
\mathcal{N}_0(r_0, s_0) &=& 1 - 2 \cdot 0 = 1 \\
\mathcal{N}_0(r_1, s_1) &=& 1 - 2 \cdot 1/2 = 0 \\
\mathcal{N}_0(r_2, s_2) &=& 1 - 2 \cdot 1/2 = 0 \\
\mathcal{N}_1(r_0, s_0) &=& 1 - 2(1 - 1/2 - 0) = 0 \\
\mathcal{N}_1(r_1, s_1) &=& 1 - 2(1 - 1/2 - 1/2) = 1 \\
\mathcal{N}_1(r_2, s_2) &=& 1 - 2(1 - 0 - 1/2) = 0 \\
\mathcal{N}_2(r_0, s_0) &=& 1 - 2 \cdot 1/2 = 0 \\
\mathcal{N}_2(r_1, s_1) &=& 1 - 2 \cdot 1/2 = 0 \\
\mathcal{N}_2(r_2, s_2) &=& 1 - 2 \cdot 0 = 1
\end{array}
$$

Automatically,

$$
\begin{array}{rcl}
\partial_r \mathcal{N}_0(r, s) &=& 0 \\
\partial_r \mathcal{N}_1(r, s) &=& 2 \\
\partial_r \mathcal{N}_2(r, s) &=& -2 \\
\partial_s \mathcal{N}_0(r, s) &=& -2 \\
\partial_s \mathcal{N}_1(r, s) &=& 2 \\
\partial_s \mathcal{N}_2(r, s) &=& 0
\end{array}
$$

Another way to obtain the basis functions is as follows.

$$
f^h(r, s) = a + br + cs
$$

$$
\begin{array}{rcl}
f_0 &=& f^h(r_0, s_0) = a + b/2 \\
f_1 &=& f^h(r_1, s_1) = a + b/2 + c/2 \\
f_2 &=& f^h(r_2, s_2) = a + c/2
\end{array}
$$

or,

$$
\begin{pmatrix} 1 & 1/2 & 0 \\ 1 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix}
$$

i.e.

$$
\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 2 & -2 \\ -2 & 2 & 0 \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix}
$$

so

$$
\begin{array}{rcl}
f^h(r, s) &=& a + br + cs \\
&=& (f_0 - f_1 + f_2) + 2(f_1 - f_2)r + 2(-f_0 + f_1)s \\
&=& (1 - 2s)f_0 + (-1 + 2r + 2s)f_1 + (1 - 2r)f_2 \\
&=& \displaystyle\sum_{i=0}^{2} \mathcal{N}_i(r, s) f_i
\end{array}
$$

8.4 Elements and basis functions in 3D

8.4.1 Trilinear basis functions (Q_1)

Following Section 8.3.1, we can simply write:

$$
\begin{aligned}
\mathcal{N}_1(r,s,t) &= 0.125(1-r)(1-s)(1-t) \\
\mathcal{N}_2(r,s,t) &= 0.125(1+r)(1-s)(1-t) \\
\mathcal{N}_3(r,s,t) &= 0.125(1+r)(1+s)(1-t) \\
\mathcal{N}_4(r,s,t) &= 0.125(1-r)(1+s)(1-t) \\
\mathcal{N}_5(r,s,t) &= 0.125(1-r)(1-s)(1+t) \\
\mathcal{N}_6(r,s,t) &= 0.125(1+r)(1-s)(1+t) \\
\mathcal{N}_7(r,s,t) &= 0.125(1+r)(1+s)(1+t) \\
\mathcal{N}_8(r,s,t) &= 0.125(1-r)(1+s)(1+t)
\end{aligned}
$$

These basis functions are the product of the linear basis functions of Section 8.2.1.
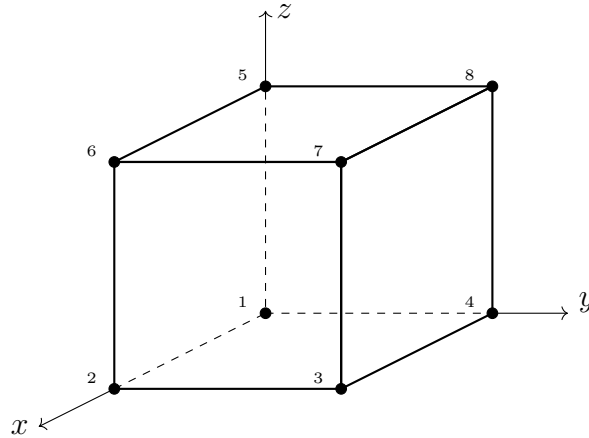
Note that the figure above depicts an element in x, y, z space, not the reference element in r, s, t space which is centered around the origin.

8.4.2 Triquadratic basis functions (Q_2)
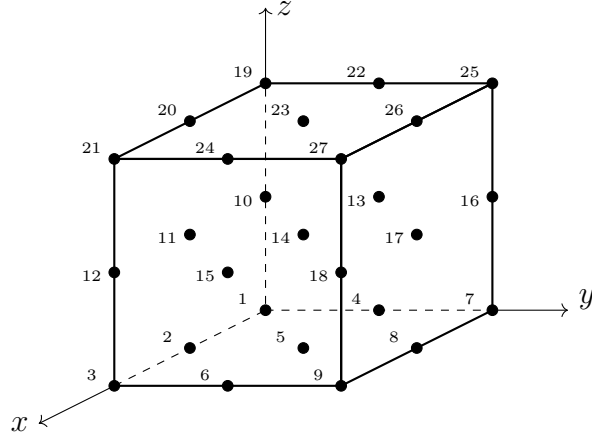
Note that the figure above depicts an element in x, y, z space, not the reference element in r, s, t space which is centered around the origin.

$$
\begin{aligned}
\mathcal{N}_1 &= 0.5r(r-1) \cdot 0.5s(s-1) \cdot 0.5t(t-1) \\
\mathcal{N}_2 &= (1-r^2) \cdot 0.5s(s-1) \cdot 0.5t(t-1) \\
\mathcal{N}_3 &= 0.5r(r+1) \cdot 0.5s(s-1) \cdot 0.5t(t-1) \\
\mathcal{N}_4 &= 0.5r(r-1) \cdot (1-s^2) \cdot 0.5t(t-1) \\
\mathcal{N}_5 &= (1-r^2) \cdot (1-s^2) \cdot 0.5t(t-1) \\
\mathcal{N}_6 &= 0.5r(r+1) \cdot (1-s^2) \cdot 0.5t(t-1) \\
\mathcal{N}_7 &= 0.5r(r-1) \cdot 0.5s(s+1) \cdot 0.5t(t-1) \\
\mathcal{N}_8 &= (1-r^2) \cdot 0.5s(s+1) \cdot 0.5t(t-1) \\
\mathcal{N}_9 &= 0.5r(r+1) \cdot 0.5s(s+1) \cdot 0.5t(t-1) \\
\mathcal{N}_{10} &= 0.5r(r-1) \cdot 0.5s(s-1) \cdot (1-t^2) \\
\mathcal{N}_{11} &= (1-r^2) \cdot 0.5s(s-1) \cdot (1-t^2) \\
\mathcal{N}_{12} &= 0.5r(r+1) \cdot 0.5s(s-1) \cdot (1-t^2) \\
\mathcal{N}_{13} &= 0.5r(r-1) \cdot (1-s^2) \cdot (1-t^2) \\
\mathcal{N}_{14} &= (1-r^2) \cdot (1-s^2) \cdot (1-t^2) \\
\mathcal{N}_{15} &= 0.5r(r+1) \cdot (1-s^2) \cdot (1-t^2) \\
\mathcal{N}_{16} &= 0.5r(r-1) \cdot 0.5s(s+1) \cdot (1-t^2) \\
\mathcal{N}_{17} &= (1-r^2) \cdot 0.5s(s+1) \cdot (1-t^2) \\
\mathcal{N}_{18} &= 0.5r(r+1) \cdot 0.5s(s+1) \cdot (1-t^2) \\
\mathcal{N}_{19} &= 0.5r(r-1) \cdot 0.5s(s-1) \cdot 0.5t(t+1) \\
\mathcal{N}_{20} &= (1-r^2) \cdot 0.5s(s-1) \cdot 0.5t(t+1) \\
\mathcal{N}_{21} &= 0.5r(r+1) \cdot 0.5s(s-1) \cdot 0.5t(t+1) \\
\mathcal{N}_{22} &= 0.5r(r-1) \cdot (1-s^2) \cdot 0.5t(t+1) \\
\mathcal{N}_{23} &= (1-r^2) \cdot (1-s^2) \cdot 0.5t(t+1) \\
\mathcal{N}_{24} &= 0.5r(r+1) \cdot (1-s^2) \cdot 0.5t(t+1) \\
\mathcal{N}_{25} &= 0.5r(r-1) \cdot 0.5s(s+1) \cdot 0.5t(t+1) \\
\mathcal{N}_{26} &= (1-r^2) \cdot 0.5s(s+1) \cdot 0.5t(t+1) \\
\mathcal{N}_{27} &= 0.5r(r+1) \cdot 0.5s(s+1) \cdot 0.5t(t+1)
\end{aligned}
$$

8.4.3 Linear basis functions in tetrahedra (P_1)

The location of the four notes are:

```
(r_0,s_0) = (0,0,0)
(r_1,s_1) = (1,0,0)
(r_2,s_2) = (0,1,0)
(r_3,s_3) = (0,0,1)
```

The basis polynomial is given by

$$f(r,s,t) = c_0 + c_1 r + c_2 s + c_3 t$$

and it needs to satisfy these four conditions:

$$
\begin{align}
f_1 &= f(r_1, s_1, t_1) = c_0 \tag{8.167}\\
f_2 &= f(r_2, s_2, t_2) = c_0 + c_1 \tag{8.168}\\
f_3 &= f(r_3, s_3, t_3) = c_0 + c_2 \tag{8.169}\\
f_4 &= f(r_4, s_4, t_4) = c_0 + c_3 \tag{8.170}
\end{align}
$$

which yields:

$$c_0 = f_1 \qquad c_1 = f_2 - f_1 \qquad c_2 = f_3 - f_1 \qquad c_3 = f_4 - f_1$$

Finally,

$$
\begin{align}
f(r,s,t) &= c_0 + c_1 r + c_2 s + c_3 t \\
&= f_1 + (f_2 - f_1)r + (f_3 - f_1)s + (f_4 - f_1)t \\
&= f_1(1 - r - s - t) + f_2 r + f_3 s + f_4 t \\
&= \sum_i \mathcal{N}_i(r,s,t) f_i
\end{align}
$$

Finally,

$$
\begin{align}
\mathcal{N}_1(r,s,t) &= 1 - r - s - t \\
\mathcal{N}_2(r,s,t) &= r \\
\mathcal{N}_3(r,s,t) &= s \\
\mathcal{N}_4(r,s,t) &= t
\end{align}
$$

Derivatives are trivial to obtain.

8.4.4 Enriched linear in tetrahedra(P_1^+)

These basis functions would be used in the MINI element, see Section 10.5.7.
In 3D the bubble function lools like $rst(1 - r - s - t)$ so that

$$f(r,s,t) = a + b\,r + c\,s + d\,t + e\,rst(1 - r - s - t)$$

We have node 1 at location $(r, s, t) = (0, 0, 0)$, node 2 at $(r, s, t) = (1, 0, 0)$, node 3 at $(r, s, t) = (0, 1, 0)$, node 4 at $(r, s, t) = (0, 0, 1)$ and we set the location of the bubble (node 5) at $r = s = t = 1/4$ so that

$$
\begin{aligned}
f(r_1, s_1, t_1) &= f_1 = a + b\, r_1 + c\, s_1 + d\, t_1 + e\, r_1 s_1 t_1 (1 - r_1 - s_1 - t_1) \\
f(r_2, s_2, t_2) &= f_2 = a + b\, r_2 + c\, s_2 + d\, t_2 + e\, r_2 s_2 t_2 (1 - r_2 - s_2 - t_2) \\
f(r_3, s_3, t_3) &= f_3 = a + b\, r_3 + c\, s_3 + d\, t_3 + e\, r_3 s_3 t_3 (1 - r_3 - s_3 - t_3) \\
f(r_4, s_4, t_4) &= f_4 = a + b\, r_4 + c\, s_4 + d\, t_4 + e\, r_4 s_4 t_4 (1 - r_4 - s_4 - t_4) \\
f(r_5, s_5, t_5) &= f_5 = a + b\, r_5 + c\, s_5 + d\, t_5 + e\, r_5 s_5 t_5 (1 - r_5 - s_5 - t_5) \quad (8.171)
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
f_1 &= a \\
f_2 &= a + b \\
f_3 &= a + c \\
f_4 &= a + d \\
f_5 &= a + b/4 + c/4 + d/4 + e/64(1 - 1/4 - 1/4 - 1/4) \\
    &= a + b/4 + c/4 + d/4 + e/256
\end{aligned}
$$

Then

$$
\begin{aligned}
a &= f_1 \\
b &= f_2 - f_1 \\
c &= f_3 - f_1 \\
d &= f_4 - f_1 \\
e &= 256(f_5 - a - b/4 - c/4 - d/4) \\
  &= 256(f_5 - f_1 - (f_2 - f_1)/4 - (f_3 - f_1)/4 - (f_4 - f_1)/4) \\
  &= 256(-f_1/4 - f_2/4 - f_3/4 - f_4/4 + f_5) \\
  &= 64(-f_1 - f_2 - f_3 - f_4 + 4f_5) \quad (8.172)
\end{aligned}
$$

Finally:

$$
\begin{aligned}
f(r, s, t) &= a + br + cs + dt + erst(1 - r - s - t) \\
&= f_1 + (f_2 - f_1)r + (f_3 - f_1)s + (f_4 - f_1)t + 64(-f_1 - f_2 - f_3 - f_4 + 4f_5)rst(1 - r - s - t) \\
&= f_1[1 - r - s - t - 64rst(1 - r - s - t)] \\
&+ f_2[r - 64rst(1 - r - s - t)] \\
&+ f_3[s - 64rst(1 - r - s - t)] \\
&+ f_4[t - 64rst(1 - r - s - t)] \\
&+ f_5[256rst(1 - r - s - t)] \\
&= \sum_{i=1}^{5} \mathcal{N}_i(r, s, t) f_i \quad (8.173)
\end{aligned}
$$

with

$$
\begin{array}{rcl}
\mathcal{N}_1(r,s,t) &=& 1 - r - s - t - 64rst(1 - r - s - t) \quad\quad (8.174)\\
\mathcal{N}_2(r,s,t) &=& r - 64rst(1 - r - s - t) \quad\quad (8.175)\\
\mathcal{N}_3(r,s,t) &=& s - 64rst(1 - r - s - t) \quad\quad (8.176)\\
\mathcal{N}_4(r,s,t) &=& t - 64rst(1 - r - s - t) \quad\quad (8.177)\\
\mathcal{N}_5(r,s,t) &=& +256rst(1 - r - s - t) \quad\quad (8.178)
\end{array}
$$

The derivatives are given by:

$$
\begin{array}{rcl}
\dfrac{\partial \mathcal{N}_1}{\partial r}(r,s,t) &=& -1 - 64st(1 - 2r - s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_2}{\partial r}(r,s,t) &=& +1 - 64st(1 - 2r - s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_3}{\partial r}(r,s,t) &=& -64st(1 - 2r - s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_4}{\partial r}(r,s,t) &=& -64st(1 - 2r - s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_5}{\partial r}(r,s,t) &=& 256st(1 - 2r - s - t)
\end{array}
$$

$$
\begin{array}{rcl}
\dfrac{\partial \mathcal{N}_1}{\partial s}(r,s,t) &=& -1 - 64rt(1 - r - 2s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_2}{\partial s}(r,s,t) &=& -64rt(1 - r - 2s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_3}{\partial s}(r,s,t) &=& +1 - 64rt(1 - r - 2s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_4}{\partial s}(r,s,t) &=& -64rt(1 - r - 2s - t)\\[2mm]
\dfrac{\partial \mathcal{N}_5}{\partial s}(r,s,t) &=& 256rt(1 - r - 2s - t)
\end{array}
$$

$$
\begin{array}{rcl}
\dfrac{\partial \mathcal{N}_1}{\partial t}(r,s,t) &=& -1 - 64rs(1 - r - s - 2t)\\[2mm]
\dfrac{\partial \mathcal{N}_2}{\partial t}(r,s,t) &=& -64rs(1 - r - s - 2t)\\[2mm]
\dfrac{\partial \mathcal{N}_3}{\partial t}(r,s,t) &=& -64rs(1 - r - s - 2t)\\[2mm]
\dfrac{\partial \mathcal{N}_4}{\partial t}(r,s,t) &=& +1 - 64rs(1 - r - s - 2t)\\[2mm]
\dfrac{\partial \mathcal{N}_5}{\partial t}(r,s,t) &=& 256rs(1 - r - s - 2t)
\end{array}
$$

8.4.5 Enriched quadratic basis functions in tetrahedra (P_2^+)

basis_p2p_3D.tex
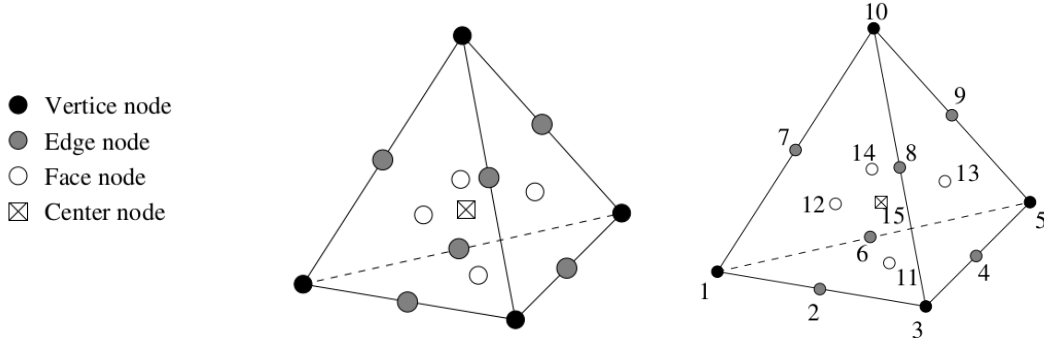
The velocity basis functions are:

$$\phi_i = \lambda_i(2\lambda_i - 1) + 3(\lambda_i\lambda_j\lambda_k + \lambda_i\lambda_j\lambda_l + \lambda_i\lambda_k\lambda_l) - 4\lambda_i\lambda_j\lambda_k\lambda_l \tag{8.179}$$

$$\phi_{ij} = 4\lambda_i\lambda_j - 12(\lambda_i\lambda_j\lambda_k + \lambda_i\lambda_j\lambda_l) + 32\lambda_i\lambda_j\lambda_k\lambda_l \tag{8.180}$$

$$\phi_{ijk} = 27\lambda_i\lambda_j\lambda_k - 108\lambda_i\lambda_j\lambda_k\lambda_l \tag{8.181}$$

$$\phi_c = 256\lambda_i\lambda_j\lambda_k\lambda_l \tag{8.182}$$

> REFS ??? better definition of functions !

8.4.6 Linear basis functions for hexahedra (P_{-1})

This is the $\boldsymbol{Q}_2 \times P_{-1}$ element. I choose the reduced coordinates of the pressure nodes to be :

point	r	s	t
1	1/2	-1/2	-1/2
2	-1/2	1/2	-1/2
3	-1/2	-1/2	1/2
4	1/2	1/2	1/2

Inside the element the pressure is given as a linear function of the reduced coordinates r, s, t:

$$p(r, s, t) = a + br + cs + dt$$

This expression must exactly interpolate the pressure at all four pressure nodes:

$$p_1 = p(r_1, s_1, t_1) = a + br_1 + cs_1 + dt_1 = a + b/2 - c/2 - d/2$$

$$p_2 = p(r_2, s_2, t_2) = a + br_2 + cs_2 + dt_2 = a - b/2 + c/2 - d/2$$

$$p_3 = p(r_3, s_3, t_3) = a + br_3 + cs_3 + dt_3 = a - b/2 - c/2 + d/2$$

$$p_4 = p(r_4, s_4, t_4) = a + br_4 + cs_4 + dt_4 = a + b/2 + c/2 + d/2$$

or,

$$\begin{pmatrix} 1 & 1/2 & -1/2 & -1/2 \\ 1 & -1/2 & +1/2 & -1/2 \\ 1 & -1/2 & -1/2 & +1/2 \\ 1 & 1/2 & +1/2 & +1/2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

The matrix is invertible and we get:

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & -1/2 & -1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & 1/2 \\ -1/2 & -1/2 & 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

323

so

$$
\begin{aligned}
p(r,s,t) &= a + br + cs + dt \\
&= \frac{1}{4}(p_1 + p_2 + p_3 + p_4) + \frac{1}{2}(p_1 - p_2 - p_3 + p_4)r + \frac{1}{2}(-p_1 + p_2 - p_3 + p_4)s + \frac{1}{2}(-p_1 - p_2 + p_3 + p_4) \\
&= \frac{1}{4}(1 + 2r - 2s - 2t)p_1 + \frac{1}{4}(1 - 2r + 2s - 2t)p_2 + \frac{1}{4}(1 - 2r - 2s + 2t)p_3 + \frac{1}{4}(1 + 2r + 2s + 2t)p_4 \\
&= \sum_{i=1}^{4} N_i(r,s,t)p_i && (8.18\text{?})
\end{aligned}
$$

with

$$
\begin{aligned}
N_1(r,s,t) &= \frac{1}{4}(1 + 2r - 2s - 2t) \\
N_2(r,s,t) &= \frac{1}{4}(1 - 2r + 2s - 2t) \\
N_3(r,s,t) &= \frac{1}{4}(1 - 2r - 2s + 2t) \\
N_4(r,s,t) &= \frac{1}{4}(1 + 2r + 2s + 2t)
\end{aligned}
$$

I could also have chosen

point	r	s	t
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

This expression must exactly interpolate the pressure at all four pressure nodes:

$$
\begin{aligned}
p_1 &= p(r_1, s_1, t_1) = a + br_1 + cs_1 + dt_1 = a \\
p_2 &= p(r_2, s_2, t_2) = a + br_2 + cs_2 + dt_2 = a + b \\
p_3 &= p(r_3, s_3, t_3) = a + br_3 + cs_3 + dt_3 = a + c \\
p_4 &= p(r_4, s_4, t_4) = a + br_4 + cs_4 + dt_4 = a + d
\end{aligned}
$$

i.e.

$$
a = p_1 \qquad b = p_2 - p1 \qquad c = p_3 - p1 \qquad d = p_4 - p1
$$

or,

$$
p^h(r,s) = a + br + cs + dt = p_1 + (p_2 - p_1)r + (p_3 - p_1)r + (p_4 - p_1)t = p_1(1 - r - s - t) + rp_2 + sp_3 + tp_4
$$

so

$$
\begin{aligned}
N_1(r,s,t) &= 1 - r - s - t && (8.184) \\
N_2(r,s,t) &= r && (8.185) \\
N_3(r,s,t) &= s && (8.186) \\
N_4(r,s,t) &= t && (8.187)
\end{aligned}
$$

8.4.7   20-node serendipity basis functions in 3D ($Q_2^{(20)}$)

The serendipity elements are those rectangular elements which have no interior nodes [1118, p91].

```
   t
   |
    .--s
   /
  r
```

```
                                    05=====20=====08
                                    |             |
                                    |             |
                17 - - - - - -19  13            16
                 .           .  |             |
                 .           .  |             |
06=====18=====07  .          .  01=====12=====04 @ r=-1
|             |   .          .
|             |   .          .
14            15  09 - - - - - -11 @ r=0
|             |
|             |
02=====10=====03 @ r=+1
```
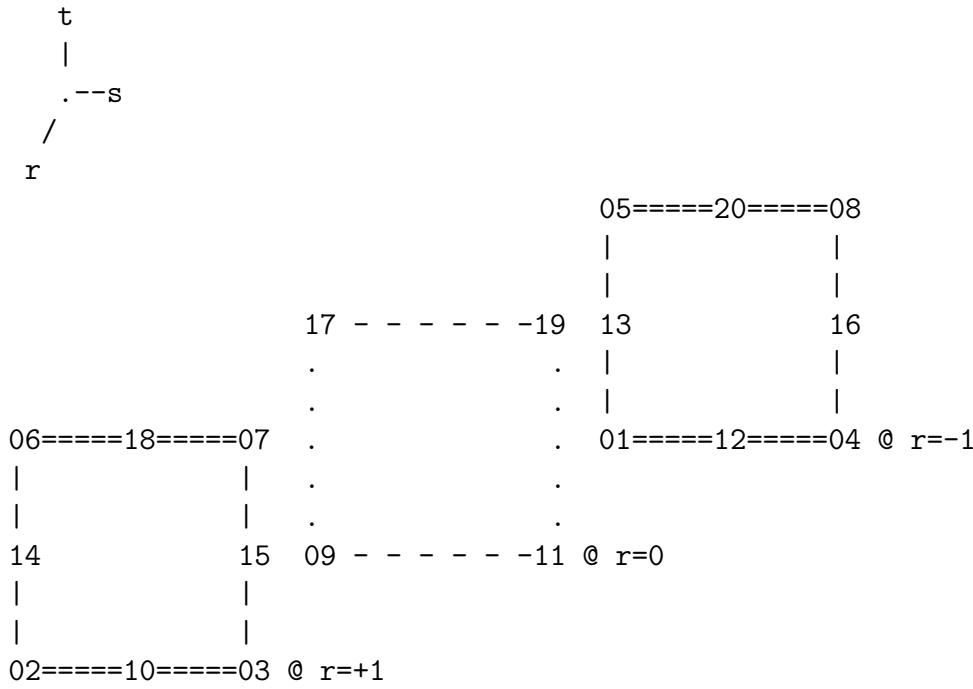
<div style="background-color:orange">find/build basis functions!</div>

8.4.8   The rotated Q_1

The nodes are not on the corners of the element but in the middle of the element faces:
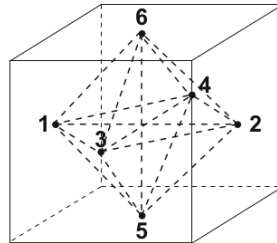
Node numbering and connectivity pattern of the reference element. Taken from [473]

We have $\tilde{Q}_1 = span\{1, r, s, t, r^2 - s^2, s^2 - t^2\}$.

The Middle Point (MP) variant   .

The basis functions are given by (see Georgiev *et al.* (2008) [473]):

$$N_1(r, s, t) = \frac{1}{6}(1 - 3r + 2r^2 - s^2 - t^2) \tag{8.188}$$

$$N_2(r, s, t) = \frac{1}{6}(1 + 3r + 2r^2 - s^2 - t^2) \tag{8.189}$$

$$N_3(r, s, t) = \frac{1}{6}(1 - r^2 - 3s + 2s^2 - t^2) \tag{8.190}$$

$$N_4(r, s, t) = \frac{1}{6}(1 - r^2 + 3s + 2s^2 - t^2) \tag{8.191}$$

$$N_5(r, s, t) = \frac{1}{6}(1 - r^2 - s^2 - 3t + 2t^2) \tag{8.192}$$

$$N_6(r, s, t) = \frac{1}{6}(1 - r^2 - s^2 + 3t + 2t^2) \tag{8.193}$$

$$\frac{\partial N_1}{\partial r} = \frac{1}{6}(-3+4r) \tag{8.194}$$

$$\frac{\partial N_2}{\partial r} = \frac{1}{6}(3+4r) \tag{8.195}$$

$$\frac{\partial N_3}{\partial r} = \frac{1}{6}(-2r) \tag{8.196}$$

$$\frac{\partial N_4}{\partial r} = \frac{1}{6}(-2r) \tag{8.197}$$

$$\frac{\partial N_5}{\partial r} = \frac{1}{6}(-2r) \tag{8.198}$$

$$\frac{\partial N_6}{\partial r} = \frac{1}{6}(-2r) \tag{8.199}$$

$$\frac{\partial N_1}{\partial s} = \frac{1}{6}(-2s) \tag{8.200}$$

$$\frac{\partial N_2}{\partial s} = \frac{1}{6}(-2s) \tag{8.201}$$

$$\frac{\partial N_3}{\partial s} = \frac{1}{6}(-3+4s) \tag{8.202}$$

$$\frac{\partial N_4}{\partial s} = \frac{1}{6}(3+4s) \tag{8.203}$$

$$\frac{\partial N_5}{\partial s} = \frac{1}{6}(-2s) \tag{8.204}$$

$$\frac{\partial N_6}{\partial s} = \frac{1}{6}(-2s) \tag{8.205}$$

$$\frac{\partial N_1}{\partial t} = \frac{1}{6}(-2t) \tag{8.206}$$

$$\frac{\partial N_2}{\partial t} = \frac{1}{6}(-2t) \tag{8.207}$$

$$\frac{\partial N_3}{\partial t} = \frac{1}{6}(-2t) \tag{8.208}$$

$$\frac{\partial N_4}{\partial t} = \frac{1}{6}(-2t) \tag{8.209}$$

$$\frac{\partial N_5}{\partial t} = \frac{1}{6}(-3+4t) \tag{8.210}$$

$$\frac{\partial N_6}{\partial t} = \frac{1}{6}(3+4t) \tag{8.211}$$

The Mid Value (MV) variant .

$$N_1(r, s, t) = \frac{1}{12}(2 - 6r + 6r^2 - 3s^2 - 3t^2) \tag{8.212}$$

$$N_2(r, s, t) = \frac{1}{12}(2 + 6r + 6r^2 - 3s^2 - 3t^2) \tag{8.213}$$

$$N_3(r, s, t) = \frac{1}{12}(2 - 3r^2 - 6s + 6s^2 - 3t^2) \tag{8.214}$$

$$N_4(r, s, t) = \frac{1}{12}(2 - 3r^2 + 6s + 6s^2 - 3t^2) \tag{8.215}$$

$$N_5(r, s, t) = \frac{1}{12}(2 - 3r^2 - 3s^2 - 6t + 6t^2) \tag{8.216}$$

$$N_6(r, s, t) = \frac{1}{12}(2 - 3r^2 - 3s^2 + 6t + 6t^2) \tag{8.217}$$

$$\frac{\partial N_1}{\partial r} = \frac{1}{12}(-6 + 12r) = \frac{1}{2}(-1 + 2r) \tag{8.218}$$

$$\frac{\partial N_2}{\partial r} = \frac{1}{12}(6 + 12r) = \frac{1}{2}(1 + 2r) \tag{8.219}$$

$$\frac{\partial N_3}{\partial r} = \frac{1}{12}(-6r) = -\frac{1}{2}r \tag{8.220}$$

$$\frac{\partial N_4}{\partial r} = \frac{1}{12}(-6r) = -\frac{1}{2}r \tag{8.221}$$

$$\frac{\partial N_5}{\partial r} = \frac{1}{12}(-6r) = -\frac{1}{2}r \tag{8.222}$$

$$\frac{\partial N_6}{\partial r} = \frac{1}{12}(-6r) = -\frac{1}{2}r \tag{8.223}$$

$$\frac{\partial N_1}{\partial s} = \frac{1}{12}(-6s) = -\frac{1}{2}s \tag{8.224}$$

$$\frac{\partial N_2}{\partial s} = \frac{1}{12}(-6s) = -\frac{1}{2}s \tag{8.225}$$

$$\frac{\partial N_3}{\partial s} = \frac{1}{12}(-6 + 12s) = \frac{1}{2}(-1 + 2s) \tag{8.226}$$

$$\frac{\partial N_4}{\partial s} = \frac{1}{12}(6 + 12s) = \frac{1}{2}(1 + 2s) \tag{8.227}$$

$$\frac{\partial N_5}{\partial s} = \frac{1}{12}(-6s) = -\frac{1}{2}s \tag{8.228}$$

$$\frac{\partial N_6}{\partial s} = \frac{1}{12}(-6s) = -\frac{1}{2}s \tag{8.229}$$

$$\frac{\partial N_1}{\partial t} = \frac{1}{12}(-6t) = -\frac{1}{2}t \tag{8.230}$$

$$\frac{\partial N_2}{\partial t} = \frac{1}{12}(-6t) = -\frac{1}{2}t \tag{8.231}$$

$$\frac{\partial N_3}{\partial t} = \frac{1}{12}(-6t) = -\frac{1}{2}t \tag{8.232}$$

$$\frac{\partial N_4}{\partial t} = \frac{1}{12}(-6t) = -\frac{1}{2}t \tag{8.233}$$

$$\frac{\partial N_5}{\partial t} = \frac{1}{12}(-6+12t) = \frac{1}{2}(-1+2t) \tag{8.234}$$

$$\frac{\partial N_6}{\partial t} = \frac{1}{12}(6+12t) = \frac{1}{2}(1+2t) \tag{8.235}$$

8.5 Low order elements recap

Let us assume a Cartesian domain discretised in $nel_x \times nel_y$ elements in 2D and $nel_x \times nel_y \times nel_z$ elements in 3D. Focusing only on the total number of velocity dofs (the values indicated after the arrows are the limits when $nel_x = nel_y = nel_z >> 1$):

- $Q_1 \times P_0$, $Q_1 \times Q_1$

$$ndof_{2D} = 2(nel_x + 1) \cdot (nel_y + 1) \qquad\qquad \rightarrow 2 \cdot nel_x^2 \qquad\qquad (8.236)$$
$$ndof_{3D} = 3(nel_x + 1) \cdot (nel_y + 1) \cdot (nel_z + 1) \qquad \rightarrow 3 \cdot nel_x^3 \qquad\qquad (8.237)$$

- $Q_2 \times P_{-1}$, $Q_2 \times Q_1$

$$ndof_{2D} = 2(2nel_x + 1) \cdot (2nel_y + 1) \qquad \rightarrow \qquad 8 \cdot nel_x^2$$
$$ndof_{3D} = 3(2nel_x + 1) \cdot (2nel_y + 1) \cdot (2nel_z + 1) \qquad \rightarrow \qquad 24 \cdot nel_x^3$$

- $Q_1^+ \times P_0$

$$ndof_{2D} = 2(nel_x + 1) \cdot (nel_y + 1) + (nel_x + 1) \cdot nel_y + nel_x \cdot (nel_y + 1) \qquad \rightarrow \qquad 4 \cdot nel_x^2$$
$$ndof_{3D} = 3(nel_x + 1) \cdot (nel_y + 1) \cdot (nel_z + 1)$$
$$+ (nel_x + 1) \cdot nel_y \cdot nel_z + nelx \cdot (nel_y + 1) \cdot nelz + nelx \cdot nely \cdot (nel_z + 1) \quad \rightarrow \quad 6 \cdot nel_x^3$$

- $Q_1 \times Q_1 + 1$ bubble

$$ndof_{2D} = 2[(nel_x + 1) \cdot (nel_y + 1) + \cdot nelx \cdot nely] \quad \rightarrow \quad 4 \cdot nel_x^2$$

- $Q_1 \times Q_1 + 2$ bubbles

$$ndof_{3D} = 3[(nel_x + 1) \cdot (nel_y + 1) \cdot (nel_z + 1) + 2 \cdot nelx \cdot nely \cdot nelz] \quad \rightarrow \quad 9 \cdot nel_x^3$$

- Rannacher-turek or DSSY:

$$ndof_{2D} = 2[(nel_y + 1) \cdot nel_x + nel_y \cdot (nel_x + 1)] \quad \rightarrow \quad 4 \cdot nel_x^2$$
$$ndof_{3D} = 3[nel_x \cdot (nel_y + 1) \cdot nel_z + (nel_x + 1) \cdot nel_y \cdot nel_z + nel_x \cdot nel_y \cdot (nel_z + 1)] \quad \rightarrow \quad 9 \cdot nel_x^3$$

If we now assume $nel_x = nel_y = nel_z$, we can then plot the values above as a function of nel_x:
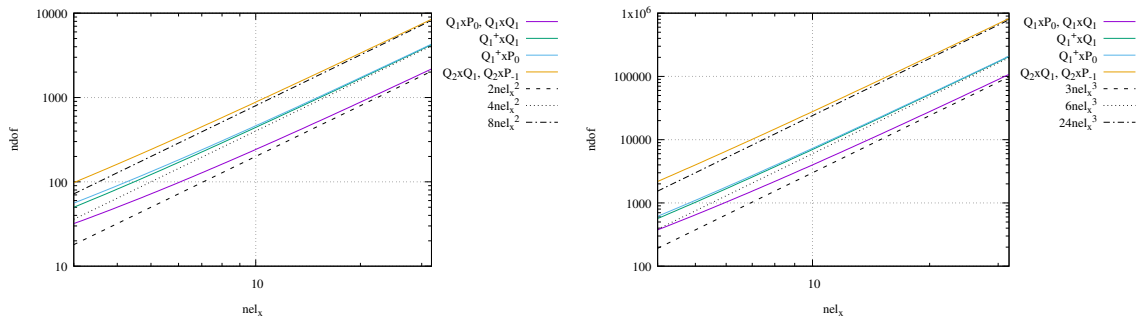
We see that the $Q_1^+ \times Q_1$ and $Q_1^+ \times P_0$ actually yield the same number of velocity dofs.

Simply based on the dof count and wishing for a (bi/tri)linear approximation for pressure, we must conclude that the $Q_1 \times Q_1+2$ bubbles is the most desirable since it is also LBB stable.

$Q_1 \times P_0, Q_1 \times Q_1$ $Q_1^+ \times Q_1$ $Q_1^+ \times P_0$ $(NC)\ Q_1 \times P_0$

$Q_2 \times Q_1$

● u,v,w   ● u,v,w on face   ● u,v,w on edge   ● u or v or w on face

$Q_1 \times P_0, Q_1 \times Q_1$ $Q_1^+ \times Q_1$?? $Q_1^+ \times P_0$

$Q_2 \times Q_1, Q_2 \times P_{-1}$ (NC) $Q_1 \times P_0$
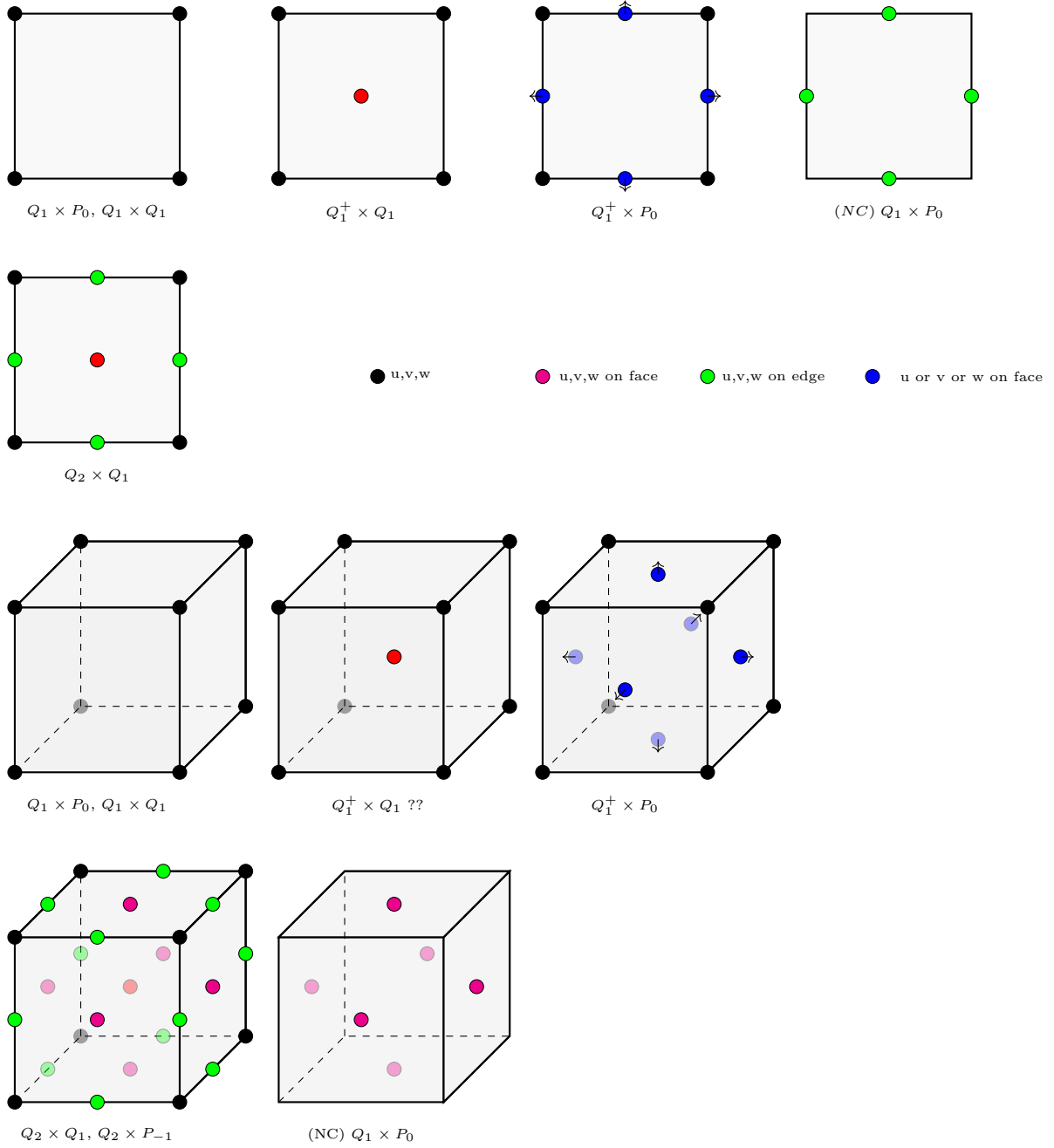
Add DSSY, RT, Q1Q1+2 bubbles

2D			
Rannacher-Turek	NCQ1P0	STONE 77	pb with buoyancy-driven flow
Lamichhane	Q1+Q1	STONE 72, STONE 74	
DSSY		STONE 77 pb with buoyancy-driven flow	
Fortin		STONE 80	

3D			
Rannacher-Turek	NCQ1P0	ELEFANT	
Lamichhane	Q1+Q1	STONE	Does not work
DSSY			
Fortin		STONE 81	

8.6   On the meaning of basis functions

In one dimension

Let us consider a 1D domain subdivided in 3 elements. We then consider the four linear basis functions attached to each node. On the following sketch these are not depicted on a single element with reduced coordinates but instead for the whole domain in the natural coordinate x:
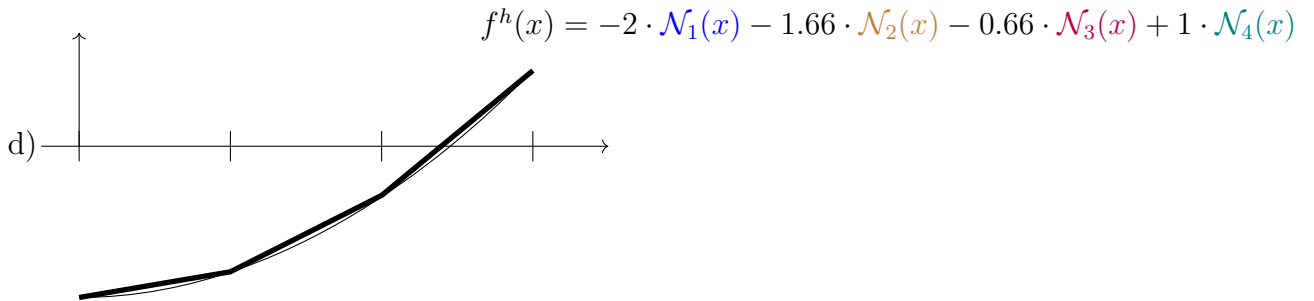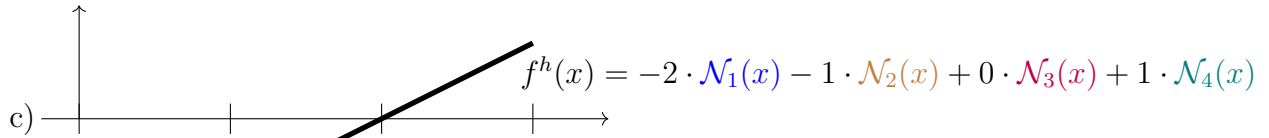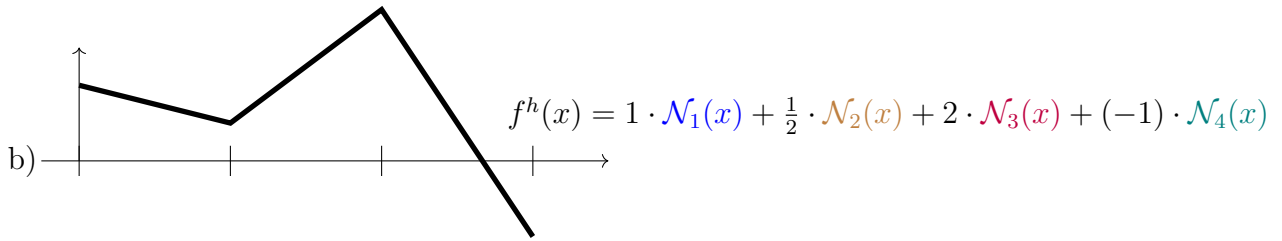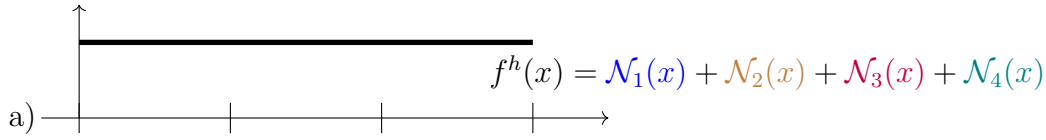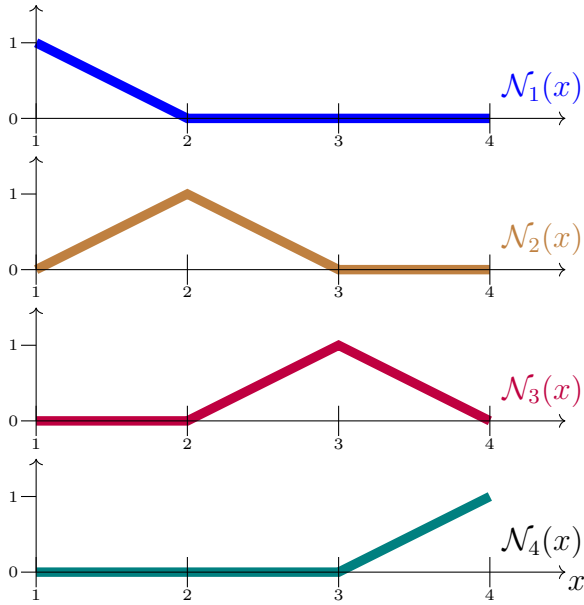
The four cases a,b,c,d are examples of combinations of these basis functions:

$$f^h(x) = \sum_{i=1}^{4} \mathcal{N}_i(x) f_i$$

Where f_i are the values associated to the four nodes. We assume that the distance h between nodes

is 1.

Example a) illustrates the fact that the sum of all basis functions must be strictly equal to one everywhere in the domain. Failing to do so would mean that the basis functions cannot represent a constant field (see Section 8.3.1).

Example b) illustrates a somewhat random combination of the basis functions, yielding a broken line.

Example c) illustrates the fact that these linear basis functions can exactly represent a linear function. When $f(x) = x - 2$, then $f_1 = f(0) = -2$, $f_2 = f(1) = -1$, $f_3 = f(2) = 0$ and $f_4 = f(3) = +1$, then $f^h(x)$ is exactly $f(x)$ on the domain.

Example d) illustrates the fact that linear basis functions cannot represent a parabola. Smaller and smaller elements will do an increasingly better job and will get closer to the curve but a systematic error will subsist.

Note that these drawings are trivial to produce since $\mathcal{N}_i(x_j) = \delta_{ij}$ by definition, so that $f^h(x_j) = f_j$.

In two dimensions

Chapter 9

Solving the heat transport equation with linear Finite Elements

9.1 The diffusion equation in 1D

Let us consider the following one-dimensional grid:
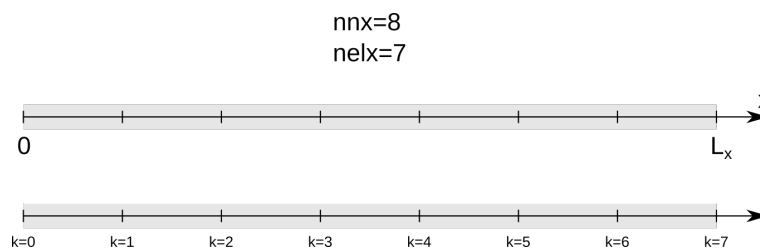
Its spans the domain Ω of length L_x. It is discretised by means of nnx nodes and $nelx = nnx - 1$ elements. Zooming in on the element e which is bounded by two nodes k and $k + 1$, its size (also sometimes called diameter) is $h_x = x_{k+1} - x_k$, and the temperature field we wish to compute is located on those nodes so that they are logically called T_k and T_{k+1}:
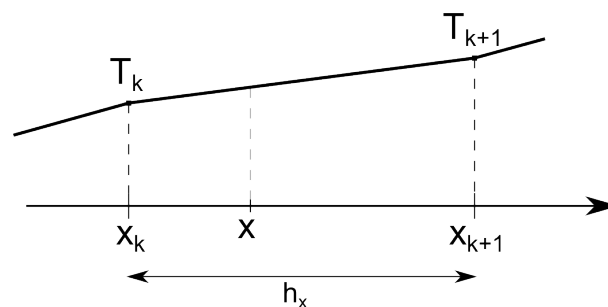
Remark. *In what follows I will indicate mathematical functions[1] by this color.*

Remark. *In what follows $\int_{\Omega_e}$ should be understood as $\int_{x_k}^{x_{k+1}}$.*

---

[1]https://en.wikipedia.org/wiki/Function_(mathematics)

From the strong form to the weak form

We focus here on the 1D diffusion equation (no advection, no heat sources, see Section 3.6):

$$\rho C_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) \tag{9.1}$$

where ρ is the density, C_p the heat capacity and k the heat conductivity. All three coefficients are assumed to be constant in space and time.

This is the strong form of the PDE to solve. I can multiply this equation by a function[2] $f(x)$ and integrate it over Ω:

$$\int_\Omega f(x)\, \rho C_p \frac{\partial T}{\partial t} dx = \int_\Omega f(x)\, \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) dx \tag{9.2}$$

Looking at the right hand side, it is of the form $\int uv'$ so that I integrate it by parts[3]:

$$\int_\Omega f(x)\, \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) dx = \left[f(x)\, k\frac{\partial T}{\partial x}\right]_{\partial\Omega} - \int_\Omega \frac{\partial f}{\partial x}\, k\frac{\partial T}{\partial x} dx \tag{9.3}$$

Assuming there is no heat flux on the boundary[4] (i.e. $q_x = -k\partial T/\partial x = 0$), then

$$\int_\Omega f(x)\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) dx = -\int_\Omega \frac{\partial f}{\partial x}k\frac{\partial T}{\partial x}dx \tag{9.4}$$

We then obtain the weak form of the diffusion equation in 1D:

$$\boxed{\int_\Omega f(x)\rho C_p\frac{\partial T}{\partial t}dx + \int_\Omega \frac{\partial f}{\partial x}k\frac{\partial T}{\partial x}dx = 0} \tag{9.5}$$

We then use the additive property of the integral $\int_\Omega \cdots = \sum\limits_{elts} \int_{\Omega_e} \cdots$ so that the equation above becomes

$$\sum_{elts}\left(\underbrace{\int_{\Omega_e} f(x)\rho C_p\frac{\partial T}{\partial t}dx}_{\Lambda_f^e} + \underbrace{\int_{\Omega_e}\frac{\partial f}{\partial x}k\frac{\partial T}{\partial x}dx}_{\Upsilon_f^e}\right) = 0 \tag{9.6}$$

From the weak form to a linear system

In order to compute these integrals (analytically or by means of a numerical quadrature), we will need to evaluate T inside the element. However, inside the element, the temperature is not known: all we hope to have at some point is the temperature at the nodes (this is what the code will compute).

For $x \in [x_k, x_{k+1}]$ we need to come up with a way to formulate the temperature in this element and we coin this T^h. It makes sense to think that $T^h(x)$ will then be a function of the temperature at the nodes, i.e. $T^h(x) = \alpha T_k + \beta T_{k+1}$ where α and β are coefficients. One over-simplified approach would be to assign $T^h(x) = (T_k + T_{k+1})/2$ (i.e. T^h is a zero-th order polynomial) but this would make the temperature discontinuous from element to element so we discard this option. A rather logical and simple solution to this problem is a linear temperature field between T_k and T_{k+1}:

---

[2]This function should be well-behaved with special properties, but we here assume it is a polynomial function.
[3]$\int_\Omega uv' = \int_{\partial\Omega} uv - \int_\Omega u'v$
[4]This is of course not always the case and we will revisit this at a later stage in the course.

$$T^h(x) = \underbrace{\frac{x_{k+1} - x}{h_x}}_{\mathcal{N}_k^\theta(x)} T_k + \underbrace{\frac{x - x_k}{h_x}}_{\mathcal{N}_{k+1}^\theta(x)} T_{k+1}$$

where $\mathcal{N}_k^\theta(x)$ is the (temperature) basis function associated to node k and $\mathcal{N}_{k+1}^\theta(x)$ is the basis function associated to node $k+1$.

Rather reassuringly, we have:

- $x = x_k$ yields $T^h(x_k) = T_k$

- $x = x_{k+1}$ yields $T^h(x_{k+1}) = T_{k+1}$

- $x = x_{1/2} = (x_k + x_{k+1})/2$ yields $T^h(x_{1/2}) = (T_k + T_{k+1})/2$

In what follows we abbreviate $\partial T^h/\partial t$ by $\dot{T}^h(x)$. Let us compute Λ_f^e and Υ_f^e separately.

$$\begin{aligned}
\Lambda_f^e &= \int_{x_k}^{x_{k+1}} f(x)\rho C_p \dot{T}^h(x) dx \\
&= \int_{x_k}^{x_{k+1}} f(x)\rho C_p \left[\mathcal{N}_k^\theta(x)\dot{T}_k + \mathcal{N}_{k+1}^\theta(x)\dot{T}_{k+1}\right] dx \\
&= \int_{x_k}^{x_{k+1}} f(x)\rho C_p \mathcal{N}_k^\theta(x)\dot{T}_k dx + \int_{x_k}^{x_{k+1}} f(x)\rho C_p \mathcal{N}_{k+1}^\theta(x)\dot{T}_{k+1} dx \\
&= \left(\int_{x_k}^{x_{k+1}} f(x)\rho C_p \mathcal{N}_k^\theta(x) dx\right)\dot{T}_k + \left(\int_{x_k}^{x_{k+1}} f(x)\rho C_p \mathcal{N}_{k+1}^\theta(x) dx\right)\dot{T}_{k+1}
\end{aligned}$$

Taking $f(x) = \mathcal{N}_k^\theta(x)$ and omitting '(x)' in the rhs:

$$\Lambda_{\mathcal{N}_k^\theta}^e = \left(\int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_k^\theta \mathcal{N}_k^\theta dx\right)\dot{T}_k + \left(\int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_k^\theta \mathcal{N}_{k+1}^\theta dx\right)\dot{T}_{k+1}$$

Taking $f(x) = \mathcal{N}_{k+1}^\theta(x)$ and omitting '(x)' in the rhs:

$$\Lambda_{\mathcal{N}_{k+1}^\theta}^e = \left(\int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_{k+1}^\theta N_k^\theta dx\right)\dot{T}_k + \left(\int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_{k+1}^\theta N_{k+1}^\theta dx\right)\dot{T}_{k+1}$$

We can rearrange these last two equations as follows:

$$\begin{pmatrix} \Lambda_{\mathcal{N}_k^\theta}^e \\ \Lambda_{\mathcal{N}_{k+1}^\theta}^e \end{pmatrix} = \begin{pmatrix} \int_{x_k}^{x_{k+1}} \mathcal{N}_k^\theta \rho C_p \mathcal{N}_k^\theta dx & \int_{x_k}^{x_{k+1}} \mathcal{N}_k^\theta \rho C_p \mathcal{N}_{k+1}^\theta dx \\ \int_{x_k}^{x_{k+1}} \mathcal{N}_{k+1}^\theta \rho C_p \mathcal{N}_k^\theta dx & \int_{x_k}^{x_{k+1}} \mathcal{N}_{k+1}^\theta \rho C_p \mathcal{N}_{k+1}^\theta dx \end{pmatrix} \cdot \begin{pmatrix} \dot{T}_k \\ \dot{T}_{k+1} \end{pmatrix}$$

and we can take the integrals outside of the matrix:

$$\begin{pmatrix} \Lambda_{\mathcal{N}_k^\theta}^e \\ \Lambda_{\mathcal{N}_{k+1}^\theta}^e \end{pmatrix} = \left[\int_{x_k}^{x_{k+1}} \rho C_p \begin{pmatrix} N_k^\theta \mathcal{N}_k^\theta & \mathcal{N}_k^\theta N_{k+1}^\theta \\ N_{k+1}^\theta \mathcal{N}_k^\theta & \mathcal{N}_{k+1}^\theta N_{k+1}^\theta \end{pmatrix} dx\right] \cdot \begin{pmatrix} \dot{T}_k \\ \dot{T}_{k+1} \end{pmatrix}$$

Finally, we can define the vectors

$$\vec{\mathcal{N}}^T = \begin{pmatrix} \mathcal{N}_k^\theta(x) \\ \mathcal{N}_{k+1}^\theta(x) \end{pmatrix}$$

337

and

$$\vec{T}^e = \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix} \qquad \dot{\vec{T}}^e = \begin{pmatrix} \dot{T}_k \\ \dot{T}_{k+1} \end{pmatrix}$$

so that

$$\begin{pmatrix} \Lambda^e_{\mathcal{N}^\theta_k} \\ \Lambda^e_{\mathcal{N}^\theta_{k+1}} \end{pmatrix} = \left(\int_{x_k}^{x_{k+1}} \vec{\mathcal{N}}^T \rho C_p \vec{\mathcal{N}} dx \right) \cdot \dot{\vec{T}}^e$$

Let us now go back to the diffusion term:

$$
\begin{aligned}
\Upsilon^e_f &= \int_{x_k}^{x^{k+1}} \frac{\partial f}{\partial x} k \frac{\partial T^h}{\partial x} dx \\
&= \int_{x_k}^{x^{k+1}} \frac{\partial f}{\partial x} k \frac{\partial(\mathcal{N}^\theta_k(x)T_k + \mathcal{N}^\theta_{k+1}(x)T_{k+1})}{\partial x} dx \\
&= \left(\int_{x_k}^{x^{k+1}} \frac{\partial f}{\partial x} k \frac{\partial \mathcal{N}^\theta_k}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x^{k+1}} \frac{\partial f}{\partial x} k \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} dx \right) T_{k+1}
\end{aligned}
$$

Taking $f(x) = \mathcal{N}^\theta_k(x)$

$$\Upsilon^e_{\mathcal{N}^\theta_k} = \left(\int_{x_k}^{x^{k+1}} k \frac{\partial \mathcal{N}^\theta_k}{\partial x} \frac{\partial \mathcal{N}^\theta_k}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x^{k+1}} k \frac{\partial \mathcal{N}^\theta_k}{\partial x} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} dx \right) T_{k+1}$$

Taking $f(x) = \mathcal{N}^\theta_{k+1}(x)$

$$\Upsilon^e_{\mathcal{N}^\theta_{k+1}} = \left(\int_{x_k}^{x^{k+1}} k \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \frac{\partial \mathcal{N}^\theta_k}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x^{k+1}} k \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} dx \right) T_{k+1}$$

$$\begin{pmatrix} \Upsilon^e_{\mathcal{N}^\theta_k} \\ \Upsilon^e_{\mathcal{N}^\theta_{k+1}} \end{pmatrix} = \begin{pmatrix} \int_{x_k}^{x^{k+1}} \frac{\partial \mathcal{N}^\theta_k}{\partial x} k \frac{\partial \mathcal{N}^\theta_k}{\partial x} dx & \int_{x_k}^{x^{k+1}} \frac{\partial \mathcal{N}^\theta_k}{\partial x} k \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} dx \\ \int_{x_k}^{x^{k+1}} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} k \frac{\partial \mathcal{N}^\theta_k}{\partial x} dx & \int_{x_k}^{x^{k+1}} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} k \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} dx \end{pmatrix} \cdot \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix}$$

or,

$$\begin{pmatrix} \Upsilon^e_{\mathcal{N}^\theta_k} \\ \Upsilon^e_{\mathcal{N}^\theta_{k+1}} \end{pmatrix} = \left[\int_{x_k}^{x^{k+1}} k \begin{pmatrix} \frac{\partial \mathcal{N}^\theta_k}{\partial x} \frac{\partial \mathcal{N}^\theta_k}{\partial x} & \frac{\partial \mathcal{N}^\theta_k}{\partial x} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \\ \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \frac{\partial \mathcal{N}^\theta_k}{\partial x} & \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \end{pmatrix} dx \right] \cdot \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix}$$

Finally, we can define the vector

$$\vec{B}^T = \begin{pmatrix} \frac{\partial \mathcal{N}^\theta_k}{\partial x} \\ \frac{\partial \mathcal{N}^\theta_{k+1}}{\partial x} \end{pmatrix}$$

so that

$$\begin{pmatrix} \Upsilon^e_{\mathcal{N}^\theta_k} \\ \Upsilon^e_{\mathcal{N}^\theta_{k+1}} \end{pmatrix} = \left(\int_{x_k}^{x_{k+1}} \vec{B}^T k \vec{B} dx \right) \cdot \vec{T}^e$$

The weak form discretised over 1 element becomes

$$\underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{\mathcal{N}}^T \rho C_p \vec{\mathcal{N}} dx \right)}_{\boldsymbol{M}^e} \cdot \dot{\vec{T}}^e + \underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{B}^T k \vec{B} dx \right)}_{\boldsymbol{K}^e_d} \cdot \vec{T}^e = 0$$

or,

$$\boxed{\boldsymbol{M}^e \cdot \dot{\vec{T}}^e + \boldsymbol{K}_d^e \cdot \vec{T}^e = 0}$$

or,

$$\boxed{\boldsymbol{M}^e \cdot \frac{\partial \vec{T}^e}{\partial t} + \boldsymbol{K}_d^e \cdot \vec{T}^e = 0}$$

$\boldsymbol{M}^e$ is commonly called the mass matrix, or capacitance matrix [1118, p103]. Note that the matrices are not coloured: the x dependence has disappeared when the integration was carried out.

In what follows I will omit the e superscript on the $\vec{T}$ term to simplify notations.

We use a first order in time discretisation for the time derivative:

$$\dot{\vec{T}} = \frac{\partial \vec{T}}{\partial t} = \frac{\vec{T}^{new} - \vec{T}^{old}}{\delta t}$$

and in the context of an implicit scheme we get

$$\boldsymbol{M}^e \cdot \frac{\vec{T}^{new} - \vec{T}^{old}}{\delta t} + \boldsymbol{K}_d^e \cdot \vec{T}^{new} = 0$$

or,

$$\boxed{(\boldsymbol{M}^e + \boldsymbol{K}_d^e \delta t) \cdot \vec{T}^{new} = \boldsymbol{M}^e \cdot \vec{T}^{old}}$$

with

$$\boldsymbol{M}^e = \int_{x_k}^{x_{k+1}} \vec{\mathcal{N}}^T \rho C_p \vec{\mathcal{N}} dx \qquad \boldsymbol{K}_d^e = \int_{x_k}^{x_{k+1}} \vec{B}^T k \vec{B} dx$$

Computing the elemental matrices

Let us compute $\boldsymbol{M}^e$ for an element:

$$\boldsymbol{M}^e = \int_{x_k}^{x_{k+1}} \vec{\mathcal{N}}^T \rho C_p \vec{\mathcal{N}} dx = \begin{pmatrix} \int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_k^\theta \mathcal{N}_k^\theta dx & \int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_k^\theta \mathcal{N}_{k+1}^\theta dx \\ \int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_{k+1}^\theta \mathcal{N}_k^\theta dx & \int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_{k+1}^\theta \mathcal{N}_{k+1}^\theta dx \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

with

$$\vec{\mathcal{N}}^T = \begin{pmatrix} \mathcal{N}_k^\theta(x) \\ \mathcal{N}_{k+1}^\theta(x) \end{pmatrix} = \begin{pmatrix} \frac{x_{k+1} - x}{h_x} \\ \frac{x - x_k}{h_x} \end{pmatrix}$$

We only need to compute 3 integrals since $M_{12} = M_{21}$. Let us start with M_{11}:

$$M_{11} = \int_{x_k}^{x_{k+1}} \rho C_p \mathcal{N}_k^\theta(x) \mathcal{N}_k^\theta(x) dx = \int_{x_k}^{x_{k+1}} \rho C_p \frac{x_{k+1} - x}{h_x} \frac{x_{k+1} - x}{h_x} dx$$

It is then customary to carry out the change of variable $x \rightarrow r$ where $r \in [-1 : 1]$ as shown hereunder:
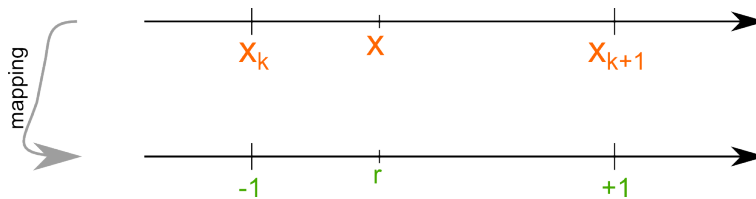
The relationships between x and r are:

$$r = \frac{2}{h_x}(x - x_k) - 1 \qquad x = \frac{h_x}{2}(1 + r) + x_k$$

In what follows we assume for simplicity that ρ and C_p are constant within each element so that:

$$M_{11} = \rho C_p \int_{x_k}^{x_{k+1}} \frac{x_{k+1} - x}{h_x} \frac{x_{k+1} - x}{h_x} dx = \frac{\rho C_p h_x}{8} \int_{-1}^{+1} (1 - r)(1 - r) dr = \rho C_p \frac{h_x}{3}$$

Similarly we arrive at

$$M_{12} = \rho C_p \int_{x_k}^{x_{k+1}} \frac{x_{k+1} - x}{h_x} \frac{x - x_k}{h_x} dx = \frac{\rho C_p h_x}{8} \int_{-1}^{+1} (1 - r)(1 + r) dr = \rho C_p \frac{h_x}{6}$$

$$M_{22} = \rho C_p \int_{x_k}^{x_{k+1}} \frac{x - x_k}{h_x} \frac{x - x_k}{h_x} dx = \frac{\rho C_p h_x}{8} \int_{-1}^{+1} (1 + r)(1 + r) dr = \rho C_p \frac{h_x}{3}$$

Finally

$$\boxed{\boldsymbol{M}^e = \rho C_p \frac{h_x}{3} \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}}$$

In the new coordinate system, the basis functions

$$\mathcal{N}_k^\theta(x) = \frac{x_{k+1} - x}{h_x} \qquad \mathcal{N}_{k+1}^\theta(x) = \frac{x - x_k}{h_x}$$

become

$$\mathcal{N}_k^\theta(r) = \frac{1}{2}(1 - r) \qquad \mathcal{N}_{k+1}^\theta(r) = \frac{1}{2}(1 + r)$$

Also,

$$\frac{\partial \mathcal{N}_k^\theta}{\partial x} = -\frac{1}{h_x} \qquad \frac{\partial \mathcal{N}_{k+1}^\theta}{\partial x} = \frac{1}{h_x}$$

so that

$$\vec{B}^T = \begin{pmatrix} \frac{\partial \mathcal{N}_k^\theta}{\partial x} \\ \frac{\partial \mathcal{N}_{k+1}^\theta}{\partial x} \end{pmatrix} = \begin{pmatrix} -\frac{1}{h_x} \\ \frac{1}{h_x} \end{pmatrix}$$

We here also assume that the heat conductivity k is constant within the element:

$$\boldsymbol{K}_d^e = \int_{x_k}^{x_{k+1}} \vec{B}^T k \vec{B} dx = k \int_{x_k}^{x_{k+1}} \vec{B}^T \vec{B} dx$$

simply becomes

$$\boldsymbol{K}_d^e = k \int_{x_k}^{x_{k+1}} \frac{1}{h_x^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} dx$$

and then

$$\boxed{\boldsymbol{K}_d^e = \frac{k}{h_x} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}}$$

In practice

Let us consider this very simple grid consisting of 4 elements/5 nodes[5]:
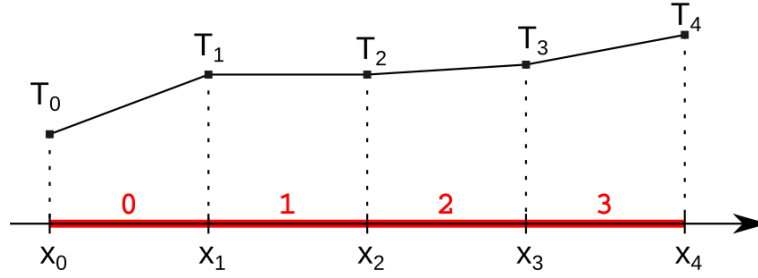
For each element we have

$$\underbrace{(\boldsymbol{M}^e + \boldsymbol{K}_d^e \, \delta t)}_{\boldsymbol{A}^e} \cdot \vec{T}^{new} = \underbrace{\boldsymbol{M}^e \cdot \vec{T}^{old}}_{\vec{b}^e}$$

with

$$\vec{T}^{new} = \begin{pmatrix} T_k^{new} \\ T_{k+1}^{new} \end{pmatrix}$$

We can write this equation very explicitly for each element:

- element 0

$$\boldsymbol{A}^0 \cdot \begin{pmatrix} T_0^{new} \\ T_1^{new} \end{pmatrix} = \vec{b}^0 \qquad \rightarrow \qquad \begin{cases} A_{00}^0 T_0^{new} + A_{01}^0 T_1^{new} = b_0^0 \\ A_{10}^0 T_0^{new} + A_{11}^0 T_1^{new} = b_1^0 \end{cases}$$

  Element 0 is made of nodes 0 and 1, so it will contributes to lines and columns 0,1 of the FE matrix:

$$\begin{pmatrix} A_{00}^0 & A_{01}^0 & 0 & 0 & 0 \\ A_{10}^0 & A_{11}^0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} T_0^{new} \\ T_1^{new} \\ T_2^{new} \\ T_3^{new} \\ T_4^{new} \end{pmatrix} = \begin{pmatrix} b_0^0 \\ b_1^0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

- element 1

$$\boldsymbol{A}^1 \cdot \begin{pmatrix} T_1^{new} \\ T_2^{new} \end{pmatrix} = \vec{b}^1 \qquad \rightarrow \qquad \begin{cases} A_{00}^1 T_1^{new} + A_{01}^1 T_2^{new} = b_0^1 \\ A_{10}^1 T_1^{new} + A_{11}^1 T_2^{new} = b_1^1 \end{cases}$$

  Element 1 is made of nodes 1 and 2, so it will contributes to lines and columns 1,2 of the FE matrix:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & A_{00}^1 & A_{01}^1 & 0 & 0 \\ 0 & A_{10}^1 & A_{11}^1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} T_0^{new} \\ T_1^{new} \\ T_2^{new} \\ T_3^{new} \\ T_4^{new} \end{pmatrix} = \begin{pmatrix} 0 \\ b_0^1 \\ b_1^1 \\ 0 \\ 0 \end{pmatrix}$$

- element 2

$$\boldsymbol{A}^2 \cdot \begin{pmatrix} T_2^{new} \\ T_3^{new} \end{pmatrix} = \vec{b}^2 \qquad \rightarrow \qquad \begin{cases} A_{00}^2 T_2^{new} + A_{01}^2 T_3^{new} = b_0^2 \\ A_{10}^2 T_2^{new} + A_{11}^2 T_3^{new} = b_1^2 \end{cases}$$

  Element 2 is made of nodes 2 and 3, so it will contributes to lines and columns 2,3 of the FE matrix:

---

[5]I have here adopted the illogical numbering of python – the language used by my students– so that python programmers can benchmark their code against this simple example

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{00}^2 & A_{01}^2 & 0 \\ 0 & 0 & A_{10}^2 & A_{11}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} T_0^{new} \\ T_1^{new} \\ T_2^{new} \\ T_3^{new} \\ T_4^{new} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ b_0^2 \\ b_1^2 \\ 0 \end{pmatrix}$$

- <u>element 3</u>

$$\boldsymbol{A}^3 \cdot \begin{pmatrix} T_3^{new} \\ T_4^{new} \end{pmatrix} = \vec{b}^3 \qquad \rightarrow \qquad \begin{cases} A_{00}^3 T_3^{new} + A_{01}^3 T_4^{new} = b_0^3 \\ A_{10}^3 T_3^{new} + A_{11}^3 T_4^{new} = b_1^3 \end{cases}$$

Element 3 is made of nodes 3 and 4, so it will contributes to lines and columns 3,4 of the FE matrix:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{00}^3 & A_{01}^3 \\ 0 & 0 & 0 & A_{10}^3 & A_{11}^3 \end{pmatrix} \cdot \begin{pmatrix} T_0^{new} \\ T_1^{new} \\ T_2^{new} \\ T_3^{new} \\ T_4^{new} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ b_0^3 \\ b_1^3 \end{pmatrix}$$

All equations can be cast into a single linear system: this is the assembly phase. The process can also be visualised as shown hereunder. Because nodes 2,3,4 belong to two elements elemental contributions will be summed in the matrix and the rhs:
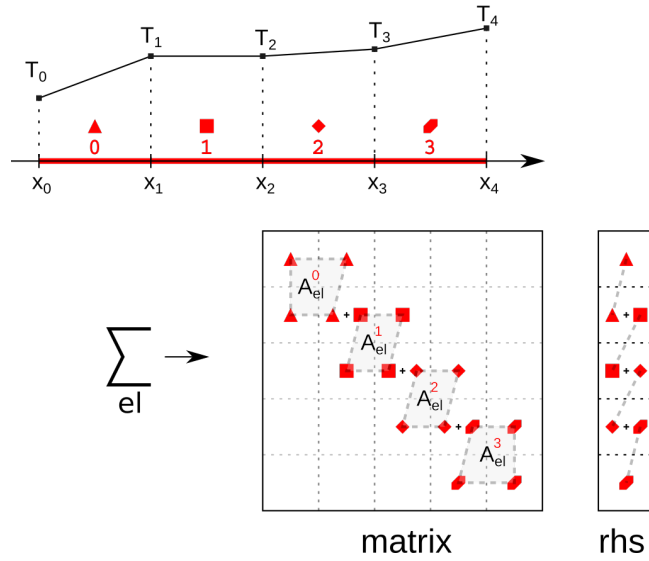
The assembled matrix and rhs are then (I have dropped the *new* superscripts):

$$\begin{pmatrix} A_{00}^0 & A_{01}^0 & 0 & 0 & 0 \\ A_{10}^0 & A_{11}^0 + A_{00}^1 & A_{01}^1 & 0 & 0 \\ 0 & A_{10}^1 & A_{11}^1 + A_{00}^2 & A_{01}^2 & 0 \\ 0 & 0 & A_{10}^2 & A_{11}^2 + A_{00}^3 & A_{01}^3 \\ 0 & 0 & 0 & A_{10}^3 & A_{11}^3 \end{pmatrix} \begin{pmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} = \begin{pmatrix} b_0^0 \\ b_1^0 + b_0^1 \\ b_1^1 + b_0^2 \\ b_1^2 + b_0^3 \\ b_1^3 \end{pmatrix}$$

Ultimately the assembled matrix system also takes the form

$$
\begin{pmatrix}
\mathcal{A}_{00} & \mathcal{A}_{01} & 0 & 0 & 0 \\
\mathcal{A}_{10} & \mathcal{A}_{11} & \mathcal{A}_{12} & 0 & 0 \\
0 & \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} & 0 \\
0 & 0 & \mathcal{A}_{32} & \mathcal{A}_{33} & \mathcal{A}_{34} \\
0 & 0 & 0 & \mathcal{A}_{43} & \mathcal{A}_{44}
\end{pmatrix}
\begin{pmatrix}
T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4
\end{pmatrix}
=
\begin{pmatrix}
b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4
\end{pmatrix}
$$

and we see that it is sparse. Its sparsity structure is easy to derive: each row corresponds to a dof, and since nodes 1 and 2 'see' each other (they belong to the same element) there will be non-zero entries in the first and second column. Likewise, node 2 'sees' node 1 (in other words, there is an edge linking nodes 1 and 2), itself, and node 3, so that there are non-zero entries in the second row at columns 1, 2, and 3.

Before we solve the system, we need to take care of boundary conditions. Let us assume that we wish to fix the temperature at node 2, or in other words we wish to set

$$
T_2 = T^o
$$

This equation can be cast as

$$
\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \end{pmatrix}
\begin{pmatrix}
T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5
\end{pmatrix}
=
\begin{pmatrix}
0 \\ T^o \\ 0 \\ 0 \\ 0
\end{pmatrix}
$$

This replaces the second line in the previous matrix equation:

$$
\begin{pmatrix}
\mathcal{A}_{11} & \mathcal{A}_{12} & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & \mathcal{A}_{32} & \mathcal{A}_{33} & \mathcal{A}_{34} & 0 \\
0 & 0 & \mathcal{A}_{43} & \mathcal{A}_{44} & \mathcal{A}_{45} \\
0 & 0 & 0 & \mathcal{A}_{54} & \mathcal{A}_{55}
\end{pmatrix}
\begin{pmatrix}
T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5
\end{pmatrix}
=
\begin{pmatrix}
b_1 \\ T^o \\ b_3 \\ b_4 \\ b_5
\end{pmatrix}
$$

That's it, we have a linear system of equations which can be solved!

The following figure presents a hand-drawn template of how a typical 1D FE code is structured:

$$\rho C_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k \frac{\partial T}{\partial x}\right)$$

transient
↳ for loop time

T , Told
= arrays (nnx)

A = FE matrix (nnx, nnx)
b = FE rhs (nnx)

$T_{old} = T_0(x)$
initial
step.

each element : $\Pi^e = \rho C_p \frac{h}{3}\begin{pmatrix}1 & \frac{1}{2} \\ \frac{1}{2} & 1\end{pmatrix}$
$K_d^e = \frac{k}{h}\begin{pmatrix}1 & -1 \\ -1 & 1\end{pmatrix}$

$h = x_{right} - x_{left}$

nelx element
nnx nodes = nelx + 1

time loop

loop over elements (0, nel)
- Π^e, K_d^e
- $A^e = \Pi^e + K_d^e \cdot \delta t$
- $\hat{b}^e = \Pi^e \cdot \vec{T}_{old}^e$
- assemble A^e, b^e ($\sum_e$)

$\begin{pmatrix} T_{old}^{left} \\ T_{old}^{right} \end{pmatrix}$

$\bar{\bar{A}} = (\backslash\backslash)$ $\vec{b} = (\vdots)$
apply b.c.
Solve $\bar{\bar{A}} \cdot \vec{T} = \vec{b}$
Told ← T
is convergence reached ?

input parameters
δt
ρ
Cp
k
Lx

practical help

for e in (0, nel) :
inode - left = e
inode - right = e+1
Told [inode. left]
Told [inode - right]

"connectivity array"

elt	0	1	2
left	0	1	2
right	1	2	3

Additional comments: A and b should be zeroed at every time step. Convergence should be tested before Told received T.

redo this for 2D

Exercise FEM-01

Write a code which solves the 1D diffusion equation in time. The initial temperature field is as follows:
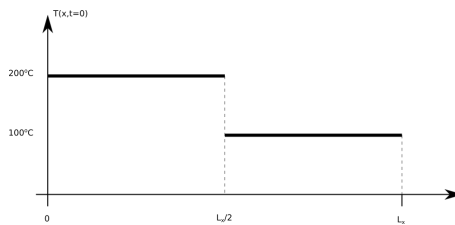
$$T(x, t = 0) = 200 \quad x < L_x/2 \qquad T(x, t = 0) = 100 \quad x \geq L_x/2$$

The domain is $L_x = 100$km and the properties of the material are $\rho = 3000 kg/m^3$, $k = 3W/m/K$, $C_p = 1000J/K$. Boundary conditions are:

$$T(t, x = 0) = 200°C \qquad T(t, x = L_x) = 100°C$$

There are `nelx` elements and `nnx` nodes. All elements are `hx` long. The code will carry out `nstep` timesteps of length `dt` with $\delta t = 0.5\frac{h_x^2}{\kappa}$.
Bonus: add a small random perturbation up to $\pm 20\%$ of h_x to each node position inside the domain and make sure that the recovered steady state solution is unchanged.

9.2 The advection-diffusion equation in 1D

We start with the 1D advection-diffusion equation

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + H \tag{9.7}$$

This is the strong form of the ODE to solve. As in the previous section, I multiply this equation by a function $f(x)$ and integrate it over the domain Ω:

$$\int_\Omega f(x) \rho C_p \frac{\partial T}{\partial t} dx + \int_\Omega f(x) \rho C_p u \frac{\partial T}{\partial x} dx = \int_\Omega f(x) \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) dx + \int_\Omega f(x) H dx$$

As in the previous section I integrate the r.h.s. by parts:

$$\int_\Omega f(x) \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) dx = \left[f(x) k \frac{\partial T}{\partial x} \right]_{\partial \Omega} - \int_\Omega \frac{\partial f}{\partial x} k \frac{\partial T}{\partial x} dx$$

Disregarding the boundary term for now, we then obtain the weak form of the diffusion equation in 1D:

$$\boxed{\int_\Omega f(x) \rho C_p \frac{\partial T}{\partial t} dx + \int_\Omega f(x) \rho C_p u \frac{\partial T}{\partial x} dx + \int_\Omega \frac{\partial f}{\partial x} k \frac{\partial T}{\partial x} dx = \int_\Omega f(x) H dx}$$

We then use the additive property of the integral $\int_\Omega \cdots = \sum_{elts} \int_{\Omega_e} \cdots$

$$\sum_{elts} \left(\underbrace{\int_{\Omega_e} f(x) \rho C_p \frac{\partial T}{\partial t} dx}_{\Lambda_f^e} + \underbrace{\int_{\Omega_e} f(x) \rho C_p u \frac{\partial T}{\partial x} dx}_{\Sigma_f^e} + \underbrace{\int_{\Omega_e} \frac{\partial f}{\partial x} k \frac{\partial T}{\partial x} dx}_{\Upsilon_f^e} - \underbrace{\int_{\Omega_e} f(x) H dx}_{\Omega_f^e} \right) = 0$$

In the element, we have seen that the temperature can be written:

$$T^h(x) = N_k^\theta(x) T_k + N_{k+1}^\theta(x) T_{k+1}$$

In the previous presentation we have computed Λ_f^e and Υ_f^e. Let us now turn to Σ_f^e and Ω_f^e.

$$
\begin{aligned}
\Sigma_f^e &= \int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial T^h}{\partial x} dx \\
&= \int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial [N_k^\theta(x) T_k + N_{k+1}^\theta(x) T_{k+1}]}{\partial x} dx \\
&= \int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial N_k^\theta}{\partial x} T_k dx + \int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial N_{k+1}^\theta}{\partial x} T_{k+1} dx \\
&= \left(\int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial N_k^\theta}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x_{k+1}} f(x) \rho C_p u \frac{\partial N_{k+1}^\theta}{\partial x} dx \right) T_{k+1}
\end{aligned}
$$

Taking $f(x) = N_k^\theta(x)$ and omitting '(x)' in the rhs:

$$\Sigma_{N_k^\theta}^e = \left(\int_{x_k}^{x_{k+1}} \rho C_p u N_k^\theta \frac{\partial N_k^\theta}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x_{k+1}} \rho C_p u N_k^\theta \frac{\partial N_{k+1}^\theta}{\partial x} dx \right) T_{k+1}$$

Taking $f(x) = N_{k+1}^\theta(x)$ and omitting '(x)' in the rhs:

$$\Sigma_{N_{k+1}^\theta}^e = \left(\int_{x_k}^{x_{k+1}} \rho C_p u N_{k+1}^\theta \frac{\partial N_k^\theta}{\partial x} dx \right) T_k + \left(\int_{x_k}^{x_{k+1}} \rho C_p u N_{k+1}^\theta \frac{\partial N_{k+1}^\theta}{\partial x} dx \right) T_{k+1}$$

$$\begin{pmatrix} \Sigma_{N_k^\theta} \\ \Sigma_{N_{k+1}^\theta} \end{pmatrix} = \begin{pmatrix} \int_{x_k}^{x_{k+1}} \rho C_p u N_k^\theta \frac{\partial N_k^\theta}{\partial x} dx & \int_{x_k}^{x_{k+1}} \rho C_p u N_k^\theta \frac{\partial N_{k+1}^\theta}{\partial x} dx \\ \int_{x_k}^{x_{k+1}} \rho C_p u N_{k+1}^\theta \frac{\partial N_k^\theta}{\partial x} dx & \int_{x_k}^{x_{k+1}} \rho C_p u N_{k+1}^\theta \frac{\partial N_{k+1}^\theta}{\partial x} dx \end{pmatrix} \cdot \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix}$$

or,

$$\begin{pmatrix} \Sigma_{N_k^\theta} \\ \Sigma_{N_{k+1}^\theta} \end{pmatrix} = \left[\int_{x_k}^{x_{k+1}} \rho C_p u \begin{pmatrix} N_k^\theta \frac{\partial N_k^\theta}{\partial x} & N_k^\theta \frac{\partial N_{k+1}^\theta}{\partial x} \\ N_{k+1}^\theta \frac{\partial N_k^\theta}{\partial x} & N_{k+1}^\theta \frac{\partial N_{k+1}^\theta}{\partial x} \end{pmatrix} dx \right] \cdot \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix}$$

Finally, we have already defined the vectors

$$\vec{N}^T = \begin{pmatrix} N_k^\theta(x) \\ N_{k+1}^\theta(x) \end{pmatrix} \qquad \vec{B}^T = \begin{pmatrix} \frac{\partial N_k^\theta}{\partial x} \\ \frac{\partial N_{k+1}^\theta}{\partial x} \end{pmatrix} \qquad \vec{T}^e = \begin{pmatrix} T_k \\ T_{k+1} \end{pmatrix}$$

so that

$$\begin{pmatrix} \Sigma_{N_k^\theta} \\ \Sigma_{N_{k+1}^\theta} \end{pmatrix} = \left(\int_{x_k}^{x_{k+1}} \vec{N}^T \rho C_p u \vec{B} dx \right) \cdot \vec{T}^e = \boldsymbol{K}_a \cdot \vec{T}^e$$

One can easily show that

$$\boldsymbol{K}_a^e = \rho C_p u \begin{pmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}$$

Note that the matrix $\boldsymbol{K}_a^e$ is *not* symmetric.

Let us now look at the source term:

$$\Omega_f^e \quad = \quad \int_{x_k}^{x^{k+1}} f(x) H(x) dx$$

Taking $f(x) = N_k^\theta(x)$:

$$\Omega_{N_k^\theta} = \int_{x_k}^{x^{k+1}} N_k^\theta(x) H(x) dx$$

Taking $f(x) = N_{k+1}^\theta(x)$:

$$\Omega_{N_{k+1}^\theta} = \int_{x_k}^{x^{k+1}} N_{k+1}^\theta(x) H(x) dx$$

We can rearrange both equations as follows:

$$\begin{pmatrix} \Omega_{N_k^\theta} \\ \Omega_{N_{k+1}^\theta} \end{pmatrix} = \begin{pmatrix} \int_{x_k}^{x^{k+1}} N_k^\theta(x) H(x) dx \\ \int_{x_k}^{x^{k+1}} N_{k+1}^\theta(x) H(x) dx \end{pmatrix}$$

or,

$$\begin{pmatrix} \Omega_{N_k^\theta} \\ \Omega_{N_{k+1}^\theta} \end{pmatrix} = \int_{x_k}^{x^{k+1}} \begin{pmatrix} N_k^\theta(x) H(x) \\ N_{k+1}^\theta(x) H(x) \end{pmatrix} dx = \left(\int_{x_k}^{x^{k+1}} \vec{N}^T H(x) dx \right)$$

The weak form discretised over 1 element becomes

$$\underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{N}^T \rho C_p \vec{N} dx \right)}_{\boldsymbol{M}^e} \cdot \dot{\vec{T}}^e + \underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{N}^T \rho C_p u \vec{B} dx \right)}_{\boldsymbol{K}_a^e} \cdot \vec{T}^e + \underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{B}^T k \vec{B} dx \right)}_{\boldsymbol{K}_d^e} \cdot \vec{T}^e = \underbrace{\left(\int_{x_k}^{x_{k+1}} \vec{N}^T H(x) dx \right)}_{\vec{F}^e}$$

or,

$$\boxed{\boldsymbol{M}^e \cdot \dot{\vec{T}}^e + (\boldsymbol{K}_d^e + \boldsymbol{K}_a^e) \cdot \vec{T}^e = \vec{F}^e}$$

or,

$$\boxed{\boldsymbol{M}^e \cdot \frac{\partial \vec{T}^e}{\partial t} + (\boldsymbol{K}_a^e + \boldsymbol{K}_d^e) \cdot \vec{T}^e = \vec{F}^e}$$

As in the diffusion case of the previous section these matrices and vectors will need to be assembled into $\boldsymbol{M}$, $\boldsymbol{K}_a$, $\boldsymbol{K}_d$, $\vec{T}$ and $\vec{F}$:

$$\boldsymbol{M} \cdot \frac{\partial \vec{T}}{\partial t} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T} = \vec{F}$$

We can revisit the time discretisation again, assuming for simplicity that the coefficients of the PDE are not time-dependent. Choosing a fully explicit approach would have us write

$$\boldsymbol{M} \cdot \frac{\vec{T}^{n+1} - \vec{T}^n}{\delta t} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T}^n = \vec{F} \quad \Rightarrow \quad \boldsymbol{M} \cdot \vec{T}^{n+1} = [\boldsymbol{M} - (\boldsymbol{K}_a + \boldsymbol{K}_d)\delta t] \cdot \vec{T}^n = \vec{F} \quad (9.8)$$

while choosing a fully implicit approach would have us write

$$\boldsymbol{M} \cdot \frac{\vec{T}^{n+1} - \vec{T}^n}{\delta t} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T}^{n+1} = \vec{F} \quad \Rightarrow \quad [\boldsymbol{M} + (\boldsymbol{K}_a + \boldsymbol{K}_d)\delta t] \cdot \vec{T}^{n+1} = \boldsymbol{M} \cdot \vec{T}^n + \vec{F}\delta t \quad (9.9)$$

We can also consider a more generic approach and write:

$$\boldsymbol{M} \cdot \frac{\vec{T}^{n+1} - \vec{T}^n}{\delta t} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot (\alpha\vec{T}^{n+1} + (1 - \alpha)\vec{T}^n) = \vec{F} \quad (9.10)$$

$$[\boldsymbol{M} + \alpha(\boldsymbol{K}_a + \boldsymbol{K}_d)\delta t] \cdot \vec{T}^{n+1} = [\boldsymbol{M} - (1 - \alpha)(\boldsymbol{K}_a + \boldsymbol{K}_d)\delta t] \cdot \vec{T}^n + \vec{F}\delta t \quad (9.11)$$

When $\alpha = 0$ we recover the explicit scheme, when $\alpha = 1$ we recover the implicit one, and when $\alpha = 1/2$ we get a so-called mid-point algorithm (Crank-Nicolson).

Write about SUPG

Exercise FEM-2

Let us consider the domain $[0, 1]$. The temperature field at $t = 0$ is given by $T = 1$ for $x < 0.25$ and $T = 0$ otherwise. The prescribed velocity is $u = 1$ and we set $nnx = 51$. Boundary conditions are $T = 1$ at $x = 0$ and $T = 0$ at $x = 1$. Only advection is present, no heat source nor diffusion.
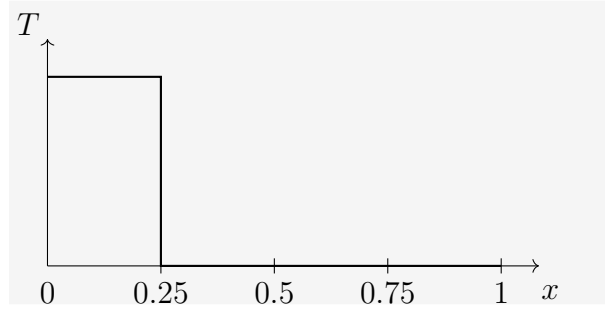
Set $\rho = C_p = 1$. Run the model for 250 time steps with $\delta t = 0.002$. Implement a fully implicit, explicit and Crank-Nicolson time discretisation.

When using Crank-Nicolson, you should then be able to recover the green line of the following figure:
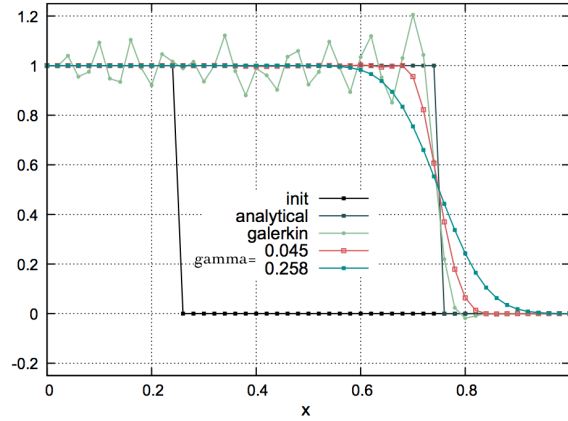
Taken from Thieulot (2011) [1343]. Note that $\tau = \gamma h / u$.

Finally, implement the SUPG method and recover the red and turquoise lines.

9.3 The advection-diffusion equation in 2D

We start from the 'bare-bones' heat transport equation (source terms are omitted):

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{\mathbf{v}} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot \left(k \vec{\nabla} T \right) \tag{9.12}$$

In what follows we assume that the velocity vield $\vec{v}$ is known so that temperature is the only unknown. Let $\mathcal{N}^\theta$ be the temperature basis functions so that the temperature inside an element is given by[6]:

$$T^h(\vec{r}) = \sum_{i=1}^{m_T} \mathcal{N}_i^\theta(\vec{r})T_i = \vec{\mathcal{N}}^\theta \cdot \vec{T} \tag{9.13}$$

where $\vec{T}$ is a vector of length m_T The weak form is then

$$\int_\Omega \mathcal{N}_i^\theta \left[\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) \right] d\Omega = \int_\Omega \mathcal{N}_i^\theta \vec{\nabla} \cdot k\vec{\nabla} T d\Omega \tag{9.14}$$

$$\underbrace{\int_\Omega \mathcal{N}_i^\theta \rho C_p \frac{\partial T}{\partial t} d\Omega}_{I} + \underbrace{\int_\Omega \mathcal{N}_i^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T d\Omega}_{II} = \underbrace{\int_\Omega \mathcal{N}_i^\theta \vec{\nabla} \cdot k\vec{\nabla} T d\Omega}_{III} \qquad i = 1, m_T \tag{9.15}$$

Looking at the first term:

$$\int_\Omega \mathcal{N}_i^\theta \rho C_p \frac{\partial T}{\partial t} d\Omega = \int_\Omega \mathcal{N}_i^\theta \rho C_p \vec{\mathcal{N}}^\theta \cdot \dot{\vec{T}} d\Omega \tag{9.16}$$

$$\tag{9.17}$$

so that when we assemble all contributions for $i = 1, m_T$ we get:

$$I = \int_\Omega \vec{\mathcal{N}}^\theta \rho C_p \vec{\mathcal{N}}^\theta \cdot \dot{\vec{T}} d\Omega = \left(\int_\Omega \rho C_p \vec{\mathcal{N}}^\theta \vec{\mathcal{N}}^\theta d\Omega \right) \cdot \dot{\vec{T}} = \boldsymbol{M}^T \cdot \dot{\vec{T}} \tag{9.18}$$

where $\boldsymbol{M}^T$ is the mass matrix of the system of size $(m_T \times m_T)$ with

$$M_{ij}^T = \int_\Omega \rho C_p \mathcal{N}_i^\theta \mathcal{N}_j^\theta d\Omega \tag{9.19}$$

Turning now to the second term:

$$\int_\Omega \mathcal{N}_i^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T d\Omega = \int_\Omega \mathcal{N}_i^\theta \rho C_p (u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y}) d\Omega \tag{9.20}$$

$$= \int_\Omega \mathcal{N}_i^\theta \rho C_p (u \frac{\partial \vec{\mathcal{N}}^\theta}{\partial x} + v \frac{\partial \vec{\mathcal{N}}^\theta}{\partial y}) \cdot \vec{T} d\Omega \tag{9.21}$$

$$\tag{9.22}$$

so that when we assemble all contributions for $i = 1, m_T$ we get:

$$II = \left(\int_\Omega \rho C_p \vec{\mathcal{N}}^\theta (u \frac{\partial \vec{\mathcal{N}}^\theta}{\partial x} + v \frac{\partial \vec{\mathcal{N}}^\theta}{\partial y}) d\Omega \right) \cdot \vec{T} = \boldsymbol{K}_a \cdot \vec{T}$$

where $\boldsymbol{K}_a$ is the advection term matrix of size $(m_T \times m_T)$ with

$$(K_a)_{ij} = \int_\Omega \rho C_p \mathcal{N}_i^\theta \left(u \frac{\partial \mathcal{N}_j^\theta}{\partial x} + v \frac{\partial \mathcal{N}_j^\theta}{\partial y} \right) d\Omega$$

---

[6]the θ superscript has been chosen to denote temperature so as to avoid confusion with the transpose operator

Now looking at the third term, we carry out an integration by part and neglect the surface term for now, so that

$$\int_\Omega \mathcal{N}_i^\theta \vec{\nabla} \cdot k \vec{\nabla} T d\Omega \;\; = \;\; -\int_\Omega k \vec{\nabla} \mathcal{N}_i^\theta \cdot \vec{\nabla} T d\Omega \tag{9.23}$$

$$= \;\; -\int_\Omega k \vec{\nabla} \mathcal{N}_i^\theta \cdot \vec{\nabla} (\vec{\mathcal{N}}^\theta \cdot \vec{T}) d\Omega \tag{9.24}$$

$$\tag{9.25}$$

with

$$\vec{\nabla} \vec{N}^\theta = \begin{pmatrix} \partial_x \mathcal{N}_1^\theta & \partial_x \mathcal{N}_2^\theta & \dots & \partial_x \mathcal{N}_{m_T}^\theta \\[2mm] \partial_y \mathcal{N}_1^\theta & \partial_y \mathcal{N}_2^\theta & \dots & \partial_y \mathcal{N}_{m_T}^\theta \end{pmatrix}$$

so that finally:

$$III = -\left(\int_\Omega k (\vec{\nabla} \vec{\mathcal{N}}^\theta)^T \cdot \vec{\nabla} \vec{\mathcal{N}}^\theta d\Omega \right) \cdot \vec{T} = -\boldsymbol{K}_d \cdot \vec{T}$$

where $\boldsymbol{K}_d$ is the diffusion term matrix:

$$\boldsymbol{K}_d = \int_\Omega k (\vec{\nabla} \vec{\mathcal{N}}^\theta)^T \cdot \vec{\nabla} \vec{\mathcal{N}}^\theta d\Omega$$

Ultimately terms I, II, III together yield:

$$\boxed{\boldsymbol{M}^\theta \cdot \dot{\vec{T}} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T} = \vec{0}}$$

add source term!!

add something about computing Kd,Ka with quadrature!

On steady states

It is said that a system is in a steady state if the (state) variables which define the behavior of the system are unchanging in time. In continuous time, this means that the partial derivative with respect to time is zero and remains so:

$$\frac{\partial}{\partial t} = 0 \qquad \forall t$$

This is irrelevant for the Stokes equations which do not contain an explicit time dependence but the heat transport equation can reach a steady state. Note that if one is only interested in the steady state solution (and not how the system gets there in time) then the heat transport equation should be solved with $\partial T / \partial t$ set to zero.

Dealing with the time discretisation

Essentially we have to solve a PDE of the type:

$$\frac{\partial T}{\partial t} = \mathcal{F}(\vec{v}, T, \vec{\nabla} T, \Delta T)$$

with $\mathcal{F} = \frac{1}{\rho C_p}(-\vec{v} \cdot \vec{\nabla} T + \vec{\nabla} \cdot k \vec{\nabla} T)$. The (explicit) forward Euler method is:

$$\frac{T^{n+1} - T^n}{\delta t} = \mathcal{F}^n(T, \vec{\nabla} T, \Delta T)$$

The (implicit) backward Euler method is:

$$\frac{T^{n+1} - T^n}{\delta t} = \mathcal{F}^{n+1}(T, \vec{\nabla} T, \Delta T)$$

and the (implicit) Crank-Nicolson algorithm is:

$$\frac{T^{n+1} - T^n}{\delta t} = \frac{1}{2}\left[\mathcal{F}^n(T, \vec{\nabla} T, \Delta T) + \mathcal{F}^{n+1}(T, \vec{\nabla} T, \Delta T)\right]$$

where the superscript n indicates the time step. The Crank-Nicolson is obviously based on the trapezoidal rule, with second-order convergence in time.

In what follows, I omit the superscript on the mass matrix to simplify notations: $\boldsymbol{M}^\theta = \boldsymbol{M}$. In terms of Finite Elements, these become:

- Explicit Forward euler:

$$\frac{1}{\delta t}(\boldsymbol{M}^{n+1} \cdot \vec{T}^{n+1} - \boldsymbol{M}^n \cdot \vec{T}^n) = -(\boldsymbol{K}_a^n + \boldsymbol{K}_d^n) \cdot \vec{T}^n$$

  or,

$$\boxed{\boldsymbol{M}^{n+1} \cdot \vec{T}^{n+1} = (\boldsymbol{M}^n - (\boldsymbol{K}_a^n + \boldsymbol{K}_d^n)\delta t) \cdot \vec{T}^n}$$

- Implicit Backward euler:

$$\frac{1}{\delta t}(\boldsymbol{M}^{n+1} \cdot \vec{T}^{n+1} - \boldsymbol{M}^n \cdot \vec{T}^n) = -(\boldsymbol{K}_a^{n+1} + \boldsymbol{K}_d^{n+1}) \cdot \vec{T}^{n+1}$$

  or,

$$\boxed{\left(\boldsymbol{M}^{n+1} + (\boldsymbol{K}_a^{n+1} + \boldsymbol{K}_d^{n+1})\delta t\right) \cdot \vec{T}^{n+1} = \boldsymbol{M}^n \cdot \vec{T}^n} \tag{9.26}$$

- Crank-Nicolson

$$\frac{1}{\delta t}\left(\boldsymbol{M}^{n+1} \cdot \vec{T}^{n+1} - \boldsymbol{M}^n \cdot \vec{T}^n\right) = \frac{1}{2}\left[-(\boldsymbol{K}_a^{n+1} + \boldsymbol{K}_d^{n+1}) \cdot \vec{T}^{n+1} - (\boldsymbol{K}_a^n + \boldsymbol{K}_d^n) \cdot \vec{T}^n\right]$$

  or,

$$\boxed{\left(\boldsymbol{M}^{n+1} + (\boldsymbol{K}_a^{n+1} + \boldsymbol{K}_d^{n+1})\frac{\delta t}{2}\right) \cdot \vec{T}^{n+1} = \left(\boldsymbol{M}^n - (\boldsymbol{K}_a^n + \boldsymbol{K}_d^n)\frac{\delta t}{2}\right) \cdot \vec{T}^n}$$

  Note that in benchmarks where the domain/grid does not deform, the coefficients do not change in space and the velocity field is constant in time, or in practice out of convenience, the $\boldsymbol{K}$ and $\boldsymbol{M}$ matrices do not change and the r.h.s. can be constructed with the same matrices as the FE matrix.

The Backward differentiation formula (see for instance Hairer & Wanner [555] or Wikipedia[7]. See also step-31 of deal.II[8]. The second-order BDF (or BDF-2) as shown in Kronbichler *et al.* (2012) [776] is as follows: it is a finite-difference quadratic interpolation approximation of the $\partial T/\partial t$ term which involves t^n, t^{n-1} and t^{n-2}:

$$\frac{\partial T}{\partial t}(t^n) \simeq \frac{1}{\delta t_n}\left(\frac{2\delta t_n + \delta t_{n-1}}{\delta t_n + \delta t_{n-1}}T^n - \frac{\delta t_n + \delta t_{n-1}}{\delta t_{n-1}}T^{n-1} + \frac{\delta t_n^2}{\delta t_{n-1}(\delta t_n + \delta t_{n-1})}T^{n-2}\right) \tag{9.27}$$

---

[7]https://en.wikipedia.org/wiki/Backward_differentiation_formula
[8]https://www.dealii.org/current/doxygen/deal.II/step_31.html

where $\delta t_n = t^n - t^{n-1}$. We also then have the approximation

$$T^n \simeq T^{n-1} + \delta t_n \frac{\partial T}{\partial t} \simeq T^{n-1} + \delta t_n \frac{T^{n-1} - T^{n-2}}{\delta t_{n-1}} = \left(1 + \frac{\delta t_n}{\delta t_{n-1}}\right) T^{n-1} + \frac{\delta t_n}{\delta t_{n-1}} T^{n-2}$$

Starting again from $\boldsymbol{M}^\theta \cdot \dot{\vec{T}} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T} = \vec{0}$, we write

$$\boldsymbol{M}^\theta \cdot \frac{1}{\delta t_n} \left(\frac{2\delta t_n + \delta t_{n-1}}{\delta t_n + \delta t_{n-1}} \vec{T}^n - \frac{\delta t_n + \delta t_{n-1}}{\delta t_{n-1}} \vec{T}^{n-1} + \frac{\delta t_n^2}{\delta t_{n-1}(\delta t_n + \delta t_{n-1})} \vec{T}^{n-2} \right) + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{T}^n = \vec{0}$$

and finally:

$$\left[\frac{2\delta t_n + \delta t_{n-1}}{\delta t_n + \delta t_{n-1}} \boldsymbol{M}^\theta + \delta t_n(\boldsymbol{K}_a + \boldsymbol{K}_d) \right] \cdot \vec{T}^n = \frac{\delta t_n + \delta t_{n-1}}{\delta t_{n-1}} \boldsymbol{M}^\theta \cdot \vec{T}^{n-1} - \frac{\delta t_n^2}{\delta_{n-1}(\delta t_n + \delta t_{n-1})} \boldsymbol{M}^\theta \cdot \vec{T}^{n-2}$$

$$(9.28)$$

For practical reasons one may wish to bring the advection term to the rhs (i.e. fully implicit) so that the matrix is symmetric. In this case the equation becomes

$$\left[\frac{2\delta t_n + \delta t_{n-1}}{\delta t_n + \delta t_{n-1}} \boldsymbol{M}^\theta + \delta t_n \boldsymbol{K}_d \right] \cdot \vec{T}^n = \frac{\delta t_n + \delta t_{n-1}}{\delta t_{n-1}} \boldsymbol{M}^\theta \cdot \vec{T}^{n-1} - \frac{\delta t_n^2}{\delta_{n-1}(\delta t_n + \delta t_{n-1})} \boldsymbol{M}^\theta \cdot \vec{T}^{n-2} - \delta t_n \boldsymbol{K}_a^\star \cdot \vec{T}^{n,\star}$$

with

$$(\cdot)^\star = \left(1 + \frac{\delta t_n}{\delta t_{n-1}}\right) (\cdot)^{n-1} + \frac{\delta t_n}{\delta t_{n-1}} (\cdot)^{n-2}$$

which denotes the extrapolation of a quantity to time n. Be aware that the $\boldsymbol{K}_a^\star$ matrix contains the velocity $\vec{v}^\star$.

Note that if all timesteps are equal, i.e. $\delta t_n = \delta t_{n-1} = \delta t$, Eq. (9.28) becomes:

$$\left[\frac{3}{2} \boldsymbol{M}^\theta + \delta t(\boldsymbol{K}_a + \boldsymbol{K}_d) \right] \cdot \vec{T}^n = \boldsymbol{M}^\theta \cdot \left(2\vec{T}^{n-1} - \frac{1}{2}\vec{T}^{n-2} \right)$$

or,

$$\left[\boldsymbol{M}^\theta + \frac{2}{3}\delta t(\boldsymbol{K}_a + \boldsymbol{K}_d) \right] \cdot \vec{T}^n = \boldsymbol{M}^\theta \cdot \left(\frac{4}{3}\vec{T}^{n-1} - \frac{1}{3}\vec{T}^{n-2} \right)$$

When the timestep δt is kept constant (which may be a bad idea with regards to the CFL condition), the backward differenciation formula family of implicit methods for the integration of ODEs are simplified. The BDF-1 is simply the backward Euler method as seen above:

$$T^{n+1} - T^n = \delta t \, \mathcal{F}^{n+1}$$

The BDF-2 is given by

$$T^{n+2} - \frac{4}{3}T^{n+1} + \frac{1}{3}T^n = \frac{2}{3}\delta t \, \mathcal{F}^{n+2}$$

In Saramito [1188] we find

$$\frac{d\phi}{dt}(t) = \frac{3\phi(t) - 4\phi(t - \delta t) + \phi(t - 2\delta t)}{2\delta t} + \mathcal{O}(\delta t^2)$$

The BDF-3 is given by

$$T^{n+3} - \frac{18}{11}T^{n+2} + \frac{9}{11}T^{n+1} - \frac{2}{11}T^n = \frac{6}{11}\delta t \, \mathcal{F}^{n+3}$$

The BDF-4 is given by

$$T^{n+4} - \frac{48}{25}T^{n+1} + \frac{36}{25}T^{n+1} - \frac{16}{25}T^{n+1} + \frac{3}{25}T^n = \frac{12}{25}\delta t \, \mathcal{F}^{n+4}$$

Each BDF-s method achieves order s.

Anisotropic heat conduction

It is most often assumed that the heat conductivity is isotropic so that one speaks of heat conductivity as a scalar k. However many materials are orthotropic and in that case the heat conductivity is a tensor $\boldsymbol{k}$ which (in 2D) writes (see Reddy [1118, p121]):

$$\boldsymbol{k} = \left(\begin{array}{cc} k_{xx} & k_{xy} \\ k_{yx} & k_{yy} \end{array} \right) = \left(\begin{array}{cc} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{array} \right) \cdot \left(\begin{array}{cc} k_1 & 0 \\ 0 & k_2 \end{array} \right) \cdot \left(\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right)$$

where k_1 and k_2 are the conductivities in the principal axes system and θ is the local orientation. In that case the diffusion term in the heat transport equation becomes $\vec{\nabla} \cdot (\boldsymbol{k} \cdot \vec{\nabla} T)$.

About the assembly

Let us consider for simplicity the following grid composed of 9 nodes and 4 Q_1 elements. Each node carries a single degree of freedom.
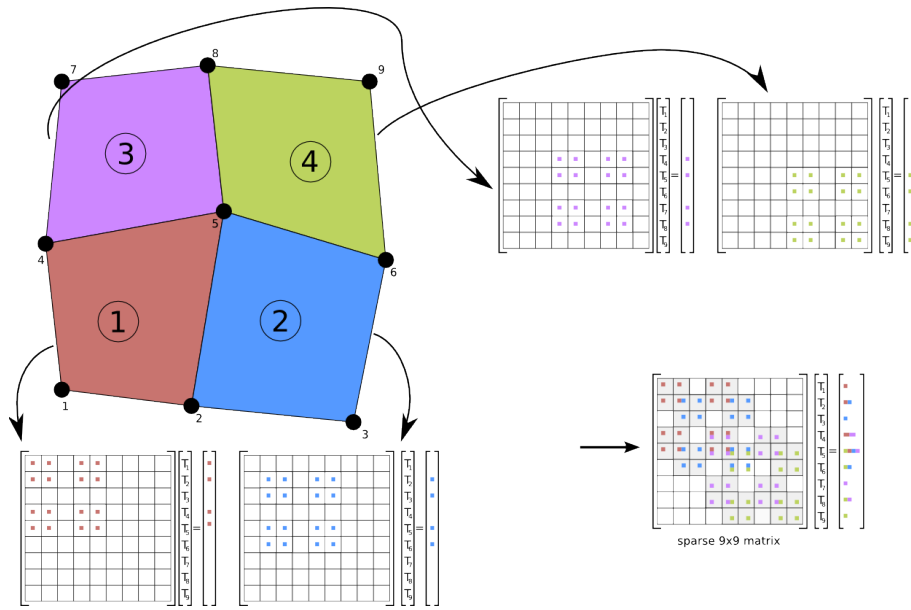
There are four elements:

- element 1 is composed of nodes $(1, 2, 5, 4) = \vec{T}^{el1}$

- element 2 is composed of nodes $(2, 3, 6, 5) = \vec{T}^{el2}$

- element 3 is composed of nodes $(4, 5, 8, 7) = \vec{T}^{el3}$

- element 4 is composed of nodes $(5, 6, 9, 8) = \vec{T}^{el4}$

For each element one has computed an elemental matrix $\boldsymbol{A}^{el}$ and a right hand side $\vec{b}^{el}$.

$$\boldsymbol{A}^{el1} \cdot \boldsymbol{T}^{el1} = \boldsymbol{b}^{el1}$$

$$\boldsymbol{A}^{el2} \cdot \boldsymbol{T}^{el2} = \boldsymbol{b}^{el2}$$

$$\boldsymbol{A}^{el3} \cdot \boldsymbol{T}^{el3} = \boldsymbol{b}^{el3}$$

$$\boldsymbol{A}^{el4} \cdot \boldsymbol{T}^{el4} = \boldsymbol{b}^{el4}$$

As seen in the 1D case, these four linear systems must be assembled in a single large matrix of size 9×9 as shown in the figure above.

9.4 Another approach to solving the advection diffusion

As we have seen above, one usually solves the heat transport equation (i.e. an advection-diffusion equation) in this form (source terms are neglected):

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot k \vec{\nabla} T \tag{9.29}$$

As we have seen in Section 3.6, the diffusion term is actually the divergence of the heat flux $\vec{q} = -k\vec{\nabla}T$. We could then choose to keep the heat flux as an unknown and solve a coupled system of equations instead:

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = -\vec{\nabla} \cdot \vec{q}$$

$$\vec{q} = -k\vec{\nabla}T$$

or,

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) + \vec{\nabla} \cdot \vec{q} = 0 \tag{9.30}$$

$$\vec{q} + k\vec{\nabla}T = \vec{0} \tag{9.31}$$

We have seen that the two left hand side terms of the first equation become $\boldsymbol{M}^\theta \cdot \dot{\vec{T}}$ and $\boldsymbol{K}_a \cdot \vec{T}$. Let $\mathcal{N}^\theta(\vec{r})$ be the temperature basis functions so that the temperature inside an element is given by

$$T^h(\vec{r}) = \sum_{i=1}^{m_T} \mathcal{N}_i^\theta(\vec{r}) T_i = \vec{\mathcal{N}}^\theta \cdot \vec{T} \tag{9.32}$$

where $\vec{T}$ is a vector of length m_T. Let $\mathcal{N}^q(\vec{r})$ be the heat flux basis functions, and let us define (in 2D) $\vec{q} = (q_x, q_y)$ so that

$$q_x^h(\vec{r}) = \sum_{i=1}^{m_q} \mathcal{N}_i^q(\vec{r}) q_{x_i} = \vec{\mathcal{N}}^q \cdot \vec{q}_x \tag{9.33}$$

$$q_y^h(\vec{r}) = \sum_{i=1}^{m_q} \mathcal{N}_i^q(\vec{r}) q_{y_i} = \vec{\mathcal{N}}^q \cdot \vec{q}_y \tag{9.34}$$

where $\vec{\mathcal{N}}^q$, $\vec{q}_x$ and $\vec{q}_y$ are vectors of length m_q. The weak form of the third term of Eq. (9.30) is then

$$\int_\Omega \mathcal{N}_i^\theta \, \nabla \cdot \vec{q} \, d\Omega = -\int_\Omega \mathcal{N}_i^\theta \left(\frac{\partial}{\partial x} q_x + \frac{\partial}{\partial y} q_y \right) d\Omega$$

$$= \int_\Omega \mathcal{N}_i^\theta \left(\frac{\partial \vec{\mathcal{N}}^q}{\partial x} \cdot \vec{q}_x + \frac{\partial \vec{\mathcal{N}}^q}{\partial y} \cdot \vec{q}_y \right) d\Omega$$

$$= \int_\Omega \mathcal{N}_i^\theta \frac{\partial \vec{\mathcal{N}}^q}{\partial x} \cdot \vec{q}_x \, d\Omega + \int_\Omega \mathcal{N}_i^\theta \frac{\partial \vec{\mathcal{N}}^q}{\partial y} \cdot \vec{q}_y \, d\Omega$$

Writing this last equation for $i = 1, ... m_T$ yields

$$\underbrace{\left(\int_\Omega \vec{\mathcal{N}}^\theta \frac{\partial \vec{\mathcal{N}}^q}{\partial x} \, d\Omega \right)}_{\boldsymbol{H}_x} \cdot \vec{q}_x + \underbrace{\left(\int_\Omega \vec{\mathcal{N}}^\theta \frac{\partial \vec{\mathcal{N}}^q}{\partial y} \, d\Omega \right)}_{\boldsymbol{H}_y} \cdot \vec{q}_y$$

In the end, we obtain:

$$\boldsymbol{M}^{\theta} \cdot \dot{\vec{T}} + \boldsymbol{K}_a \cdot \vec{T} + \boldsymbol{H}_x \cdot \vec{q}_x + \boldsymbol{H}_y \cdot \vec{q}_y = \vec{0} \tag{9.35}$$

Turning now to Eq. (9.31), its weak form is

$$\int_{\Omega} \mathcal{N}_i^q \left(\vec{q} + k \vec{\nabla} T \right) d\Omega = \vec{0}$$

and we can decompose it in its x and y components:

$$
\begin{aligned}
0 &= \int_{\Omega} \mathcal{N}_i^q \left(q_x^h + k \frac{\partial T^h}{\partial x} \right) d\Omega \\
&= \int_{\Omega} \mathcal{N}_i^q \left(\vec{\mathcal{N}}^q \cdot \vec{q}_x + k \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial x} \cdot \vec{T} \right) d\Omega \\
&= \int_{\Omega} \mathcal{N}_i^q \vec{\mathcal{N}}^q \cdot \vec{q}_x \, d\Omega + \int_{\Omega} k \mathcal{N}_i^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial x} \cdot \vec{T} \, d\Omega \\
0 &= \int_{\Omega} \mathcal{N}_i^q \left(q_y^h + k \frac{\partial T^h}{\partial y} \right) d\Omega \\
&= \int_{\Omega} \mathcal{N}_i^q \left(\vec{\mathcal{N}}^q \cdot \vec{q}_y + k \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial y} \cdot \vec{T} \right) d\Omega \\
&= \int_{\Omega} \mathcal{N}_i^q \vec{\mathcal{N}}^q \cdot \vec{q}_y \, d\Omega + \int_{\Omega} k \mathcal{N}_i^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial y} \cdot \vec{T} \, d\Omega
\end{aligned} \tag{9.36}
$$

Writing these equations for $i = 1, ... m_q$ yields:

$$
\begin{aligned}
0 &= \int_{\Omega} \vec{\mathcal{N}}^q \vec{\mathcal{N}}^q \cdot \vec{q}_x \, d\Omega + \int_{\Omega} k \vec{\mathcal{N}}^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial x} \cdot \vec{T} \, d\Omega \\
&= \underbrace{\left(\int_{\Omega} \vec{\mathcal{N}}^q \vec{\mathcal{N}}^q d\Omega \right)}_{\boldsymbol{M}^q} \cdot \vec{q}_x + \underbrace{\left(\int_{\Omega} k \vec{\mathcal{N}}^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial x} \, d\Omega \right)}_{\boldsymbol{G}_x} \cdot \vec{T}
\end{aligned} \tag{9.37}
$$

$$
\begin{aligned}
0 &= \int_{\Omega} \vec{\mathcal{N}}^q \vec{\mathcal{N}}^q \cdot \vec{q}_y \, d\Omega + \int_{\Omega} k \vec{\mathcal{N}}^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial y} \cdot \vec{T} \, d\Omega \\
&= \underbrace{\left(\int_{\Omega} \vec{\mathcal{N}}^q \vec{\mathcal{N}}^q d\Omega \right)}_{\boldsymbol{M}^q} \cdot \vec{q}_y + \underbrace{\left(\int_{\Omega} k \vec{\mathcal{N}}^q \frac{\partial \vec{\mathcal{N}}^{\theta}}{\partial y} \, d\Omega \right)}_{\boldsymbol{G}_y} \cdot \vec{T}
\end{aligned} \tag{9.38}
$$

Finally Eqs. (9.35,9.37,9.38) can be combined and yield the following system (assuming an implicit backward Euler time scheme):

$$
\begin{pmatrix} \boldsymbol{M}^{\theta} + \boldsymbol{K}_a \delta t & \boldsymbol{H}_x \delta t & \boldsymbol{H}_y \delta t \\ \boldsymbol{G}_x & \boldsymbol{M}^q & 0 \\ \boldsymbol{G}_y & 0 & \boldsymbol{M}^q \end{pmatrix} \cdot \begin{pmatrix} \vec{T}^{n+1} \\ \vec{q}_x^{n+1} \\ \vec{q}_y^{n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}^{\theta} \cdot \vec{T}^n \\ \vec{0} \\ \vec{0} \end{pmatrix}
$$

If we choose $m_q = m_T$ and $\mathcal{N}^q = \mathcal{N}^{\theta}$ then $\boldsymbol{M}^{\theta} = \boldsymbol{M}^q = \boldsymbol{M}$ so that

$$
\begin{pmatrix} \boldsymbol{M} + \boldsymbol{K}_a \delta t & \boldsymbol{H}_x \delta t & \boldsymbol{H}_y \delta t \\ \boldsymbol{G}_x & \boldsymbol{M} & 0 \\ \boldsymbol{G}_y & 0 & \boldsymbol{M} \end{pmatrix} \cdot \begin{pmatrix} \vec{T}^{n+1} \\ \vec{q}_x^{n+1} \\ \vec{q}_y^{n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M} \cdot \vec{T}^n \\ \vec{0} \\ \vec{0} \end{pmatrix}
$$

Also, if k is constant in space then $\boldsymbol{G}_{x,y} = k\boldsymbol{H}_{x,y}$. Rather interestingly, one could write Eqs. (9.37,9.38) as

$$\vec{q}^{\,n+1}_{x} = -(\boldsymbol{M}^q)^{-1} \cdot \boldsymbol{G}_x \cdot \vec{T}^{n+1} \tag{9.39}$$

$$\vec{q}^{\,n+1}_{y} = -(\boldsymbol{M}^q)^{-1} \cdot \boldsymbol{G}_y \cdot \vec{T}^{n+1} \tag{9.40}$$

and inject it in Eq. (9.35) to yield:

$$\boldsymbol{M}^\theta \cdot \dot{\vec{T}} + [\boldsymbol{K}_a - \boldsymbol{H}_x \cdot (\boldsymbol{M}^q)^{-1} \cdot \boldsymbol{G}_x - \boldsymbol{H}_y \cdot (\boldsymbol{M}^q)^{-1} \cdot \boldsymbol{G}_y] \cdot \vec{T}^{n+1} = \vec{0} \tag{9.41}$$

which means that we can directly solve for temperature! Rather interestingly, it is not equivalent to Eq. (9.26). Food for thought ...

We will see that this approach bears a lot of resemblance to the one taken in the context of Discontinuous Galerkin methods.

9.5 The advection-diffusion eq in axisymmetric cylindrical coordinates

We start from

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = k \Delta T$$

The temperature gradient in cylindrical coordinates is

$$\vec{\nabla} T = \begin{pmatrix} \partial_r T \\ \frac{1}{r} \partial_\theta T \\ \partial_z T \end{pmatrix}$$

Since $\boldsymbol{v}_\theta = 0$ and also $\partial_\theta T = 0$ then

$$\vec{v} \cdot \vec{\nabla} T = \boldsymbol{v}_r \frac{\partial T}{\partial r} + \boldsymbol{v}_z \frac{\partial T}{\partial z}$$

and we have the Laplace operator (terms in ∂_θ have been left out):

$$\Delta T = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T}{\partial r} \right) + \frac{\partial^2 T}{\partial z^2}$$

However for the FE formulation we will formulate the equation as

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot (k \vec{\nabla} T)$$

After multiplying this equation by a test function and integrating over the domain, the diffusion term is integrated by parts (surface terms are per usual discarded), and we finally obtain

$$\boldsymbol{M} \cdot \frac{\partial \vec{\mathcal{T}}}{\partial t} + (\boldsymbol{K}_a + \boldsymbol{K}_d) \cdot \vec{\mathcal{T}} = \vec{0}$$

with

$$\boldsymbol{K}_a = \int \rho C_p \vec{N}^T (\vec{\mathbf{v}} \cdot \boldsymbol{B}) \, dV \tag{9.42}$$

$$\boldsymbol{K}_d = \int k \boldsymbol{B}^T \cdot \boldsymbol{B} \, dV \tag{9.43}$$

where the matrix $\boldsymbol{B}$ is identical in this case to the 2D Cartesian one.

It looks like switching from 2D Cartesian to 3D cylindrical axisymmetric does not introduce any change in the formulation. A bit too good to be true ?

📚 Relevant Literature: check section 2.9 of J.N. Reddy and D.K. Gartling. *The Finite Element Method in Heat Transfer and Fluid Dynamics.* CRC Press, 2010. ISBN: 978-1-4200-8598-3.

9.6 The SUPG formulation for the energy equation

As abundantly documented in the literature advection needs to be stabilised as it otherwise showcases non-negligible under- and overshoots. A standard approach is the Streamline Upwind Petrov Galerkin (SUPG) method[9].

A nice overview of upwind techniques and how SUPG came to be is to be found in Chapter 2 of Donea and Huerta [363] or in Brooks and Hughes [167]. Hughes, Mallet, and Mizukami [643] (1986) present a review of the SUPG method and discuss an additional discontinuity-capturing term. So do Tezduyar and Park [1337] (1986). Note that the idea of upwind goes back to the seventies, see for example Heinrich, Huyakorn, Zienkiewicz, and Mitchell [592] (1977).

It is compared to other methods in the context of mantle convection in Malevsky & Yuen (1991) [873] (rotating cone, convection in a box). Note that SUPG is not restricted to advection(-diffusion) equations: it can also be applied to the Navier-Stokes equations[10].

As explained in the step-9 tutorial of deal.II (`https://www.dealii.org/current/doxygen/deal.II/step_9.html`), the term 'Petrov-Galerkin' can be explained as follows:

> A "Galerkin method" is one where one obtains the weak formulation by multiplying the equation by a test function v (and then integrating over Ω) where the functions v are from the same space as the solution u (though possibly with different boundary values). But this is not strictly necessary: One could also imagine choosing the test functions from a different set of functions, as long as that different set has "as many dimensions" as the original set of functions so that we end up with as many independent equations as there are degrees of freedom (where all of this needs to be appropriately defined in the infinite-dimensional case). Methods that make use of this possibility (i.e., choose the set of test functions differently than the set of solutions) are called "Petrov-Galerkin" methods.

9.6.1 The SUPG method

We start from the 'bare-bones' heat transport equation (source terms are omitted):

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{\mathbf{v}} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot (k \vec{\nabla} T) \tag{9.44}$$

---

[9]Note that Malevsky and Yuen [873] (1991) speak of "streamwise-upwinding Petrov-Galerkin" method

[10]`https://en.wikipedia.org/wiki/Streamline_upwind_Petrov-Galerkin_pressure-stabilizing_Petrov-Galerkin_formulation_for_incompressible_Navier-Stokes_equations`

which we once again write

$$\rho C_p \left(\dot{T} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot (k \vec{\nabla} T) \tag{9.45}$$

In what follows we assume that the velocity vield $\vec{v}$ is known so that temperature is the only unknown. Let $\mathcal{N}_i^\theta$ be the temperature basis function at node i so that the temperature inside an element is given by[11]:

$$T^h(\vec{r}) = \sum_{i=1}^{m_T} \mathcal{N}_i^\theta(\vec{r}) \, T_i = \vec{\mathcal{N}}^\theta \cdot \vec{T} \tag{9.46}$$

where $\vec{T}$ and $\vec{\mathcal{N}}^\theta$ are vectors of length m_T. Also, in 2d:

$$\vec{\nabla} \vec{\mathcal{N}}^\theta = \begin{pmatrix} \partial_x \mathcal{N}_1^\theta & \partial_x \mathcal{N}_2^\theta & \dots & \partial_x \mathcal{N}_{m_T}^\theta \\[2mm] \partial_y \mathcal{N}_1^\theta & \partial_y \mathcal{N}_2^\theta & \dots & \partial_y \mathcal{N}_{m_T}^\theta \end{pmatrix}$$

The weak form is then of the form

$$\int_{\Omega_e} \mathcal{N}_i^\theta \left[\rho C_p \left(\dot{T} + \vec{v} \cdot \vec{\nabla} T \right) \right] d\Omega = \int_{\Omega_e} \mathcal{N}_i^\theta \vec{\nabla} \cdot (k \vec{\nabla} T) d\Omega \tag{9.47}$$

or, rather we look for the discrete temperature T^h such that

$$\underbrace{\int_{\Omega_e} \mathcal{N}_i^\theta \rho C_p \dot{T}^h d\Omega}_{I} + \underbrace{\int_{\Omega_e} \mathcal{N}_i^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T^h d\Omega}_{II} = \underbrace{\int_{\Omega_e} \mathcal{N}_i^\theta \vec{\nabla} \cdot (k \vec{\nabla} T^h) d\Omega}_{III} \qquad i = 1, m_T$$

We have already worked out in Section 9.3 the terms I, II and III. The streamline upwind Petrov-Galerkin (SUPG) method adds the following stabilisation term to the left hand side of this equation (see Section 2.4 in Donea and Huerta [363]):

$$\int_{\Omega_e} (\vec{v} \cdot \vec{\nabla} \mathcal{N}_i^\theta) \tau \mathcal{R}(T^h, \vec{v}) \, d\Omega \tag{9.48}$$

where $\mathcal{R}$ is the residual defined as

$$\mathcal{R}(T^h, \vec{v}) = \rho C_p (\dot{T}^h + \vec{v} \cdot \vec{\nabla} T^h) - \vec{\nabla} \cdot k \vec{\nabla} T^h \tag{9.49}$$

and τ is a stabilization parameter (see Section 9.6.2).

Since I, II and III have been worked out already, we here after only focus on the additional SUPG terms.

Given Eq. (9.48) and (9.49) we then have three terms to deal with:

$$IV = \int_{\Omega_e} \tau (\vec{v} \cdot \nabla \mathcal{N}_i^\theta)(\rho C_p \dot{T}^h) d\Omega \tag{9.50}$$

$$V = \int_{\Omega_e} \tau (\vec{v} \cdot \nabla \mathcal{N}_i^\theta)(\rho C_p \vec{v} \cdot \vec{\nabla} T^h) d\Omega \tag{9.51}$$

$$VI = \int_{\Omega_e} \tau (\vec{v} \cdot \nabla \mathcal{N}_i^\theta)(-\vec{\nabla} \cdot k \vec{\nabla} T^h) d\Omega \tag{9.52}$$

---

[11]the θ superscript has been chosen to denote temperature so as to avoid confusion with the transpose operator

and we must solve $I + II + IV + V + VI = III$. We can compute the quantities $I + IV$, $II + V$ and $III - VI$:

$$
\begin{aligned}
I + IV &= \int_{\Omega_e} \mathcal{N}_i^\theta \rho C_p \dot{T}^h d\Omega + \int_{\Omega_e} (\vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta) \tau (\rho C_p \dot{T}^h) d\Omega \\
&= \int_{\Omega_e} (\mathcal{N}_i^\theta + \tau \vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta)(\rho C_p \dot{T}^h) d\Omega
\end{aligned} \tag{9.53}
$$

$$
\begin{aligned}
II + V &= \int_{\Omega_e} \mathcal{N}_i^\theta \rho C_p \vec{\mathbf{v}} \cdot \vec{\nabla} T^h d\Omega + \int_{\Omega_e} (\vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta) \tau (\rho C_p \vec{\mathbf{v}} \cdot \vec{\nabla} T^h) d\Omega \\
&= \int_{\Omega_e} (\mathcal{N}_i^\theta + \tau \vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta)(\rho C_p \vec{\mathbf{v}} \cdot \vec{\nabla} T^h) d\Omega
\end{aligned} \tag{9.54}
$$

$$
\begin{aligned}
III - VI &= \int_{\Omega_e} \mathcal{N}_i^\theta \vec{\nabla} \cdot (k \vec{\nabla} T^h) d\Omega - \int_{\Omega_e} \tau (\vec{\mathbf{v}} \cdot \nabla \mathcal{N}_i^\theta)(-\vec{\nabla} \cdot k \vec{\nabla} T^h) d\Omega \\
&= \int_{\Omega_e} (\mathcal{N}_i^\theta + \tau \vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta) \vec{\nabla} \cdot (k \vec{\nabla} T^h) d\Omega
\end{aligned} \tag{9.55}
$$

Remark. *If the equation is a pure advection equation, then the heat conductivity $k = 0$ so $III = VI = 0$.*

We see that both $I + IV$ and $II + V$ contain the term $\mathcal{N}_i^\theta + \tau \vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta$ which we can interpret as a 'modified' basis function:

$$\boxed{\underline{\mathcal{N}}_i^\theta = \mathcal{N}_i^\theta + \tau \vec{\mathbf{v}} \cdot \vec{\nabla} \mathcal{N}_i^\theta}$$

This yields the following modified elemental matrices

$$
\mathbb{M}^\theta \quad \rightarrow \quad \underline{\mathbb{M}}^\theta = \int_{\Omega_e} \rho C_p \underline{\vec{\mathcal{N}}}^\theta \vec{\mathcal{N}}^\theta \, d\Omega \tag{9.56}
$$

$$
\mathbb{K}_a \quad \rightarrow \quad \underline{\mathbb{K}}_a = \int_{\Omega_e} \rho C_p \underline{\vec{\mathcal{N}}}^\theta (\vec{\mathbf{v}} \cdot \vec{\nabla} \vec{\mathcal{N}}^\theta) \, d\Omega \tag{9.57}
$$

$$
\mathbb{K}_d \quad \rightarrow \quad \underline{\mathbb{K}}_d = \int_{\Omega_e} k \vec{\nabla} \underline{\vec{\mathcal{N}}}^\theta \cdot \vec{\nabla} \vec{\mathcal{N}}^\theta d\Omega \tag{9.58}
$$

These notations may look attractive but the modified diffusion matrix contains the gradient of the modified basis functions and working this out in practice proves to be cumbersome and not practical for higher order basis functions. Also the surface term arising from the integration by parts containing these modified basis functions might lose its physical significance. As such, we will therefore keep $\mathbb{K}_d$ (coming from the integration by parts of III, and its associated surface flux term not shown above) and VI separated.

Under the assumption that k is constant within the element Ω_e, we have

$$
VI = \int_{\Omega_e} \tau (\vec{\mathbf{v}} \cdot \nabla \mathcal{N}_i^\theta)(-k \Delta T^h) d\Omega
$$

with

$$
\Delta T^h = \Delta \left(\sum_{i=1}^{m_T} \mathcal{N}_i^\theta(\vec{r}) T_i \right) = \sum_{i=1}^{m_T} \Delta \mathcal{N}_i^\theta(\vec{r}) T_i = \Delta \vec{\mathcal{N}}^\theta \cdot \vec{T}
$$

If the basis functions are first-order ones, then $VI = 0$ since $\Delta \vec{\mathcal{N}}^\theta = \vec{0}$.

Remark. *The modified mass matrix is not symmetric anymore. This is not necessarily problematic since the advection matrix is also not symmetric so that the solver of choice for the equation already can deal with such matrices.*

The SUPG-stabilised pure advection equation is extensively tested in STONE 43.

9.6.2 About the choice of the parameter τ

1. The approach used in step-63 and subsequently in ASPECT is the one in John & Knobloch [692, 758] which is also to be found in early 80's papers [167, 638]:

$$\tau_1 = \frac{h}{2|\vec{v}|p}\left(\coth(\mathsf{Pe}) - \frac{1}{\mathsf{Pe}}\right) \qquad \mathsf{Pe} = \frac{|\vec{v}|h}{2\kappa p}$$

where τ is computed on each cell, h being its diameter in the direction of $\vec{v}$, and p is the order of the approximation (i.e. the maximum degree of polynomials).

```
template <int dim>
double compute_stabilization_delta(const double      hk,
                                   const double      eps,
                                   const Tensor<1, dim> dir,
                                   const double      pk)
{
  const double Peclet = dir.norm() * hk / (2.0 * eps * pk);
  const double coth =
    (1.0 + std::exp(-2.0 * Peclet)) / (1.0 - std::exp(-2.0 * Peclet));

  return hk / (2.0 * dir.norm() * pk) * (coth - 1.0 / Peclet);
}
```

Taken from step-63 webpage at `https://www.dealii.org/current/doxygen/deal.II/step_63.html`

A quick note about

$$\beta(x) = \coth(x) - \frac{1}{x}$$

If $x \to \infty$ then $1/x \to 0$ and then

$$\beta(x) \to \coth(x) = \frac{\exp(x) + \exp(-x)}{\exp(x) - \exp(-x)} = \frac{1 + \exp(-2x)}{1 - \exp(-2x)} \to 1$$

If $x \to 0$ then we must use the Taylor expansion of the exp function:

$$\exp(x) \simeq 1 + x + x\frac{x^2}{2} + \frac{x^3}{6}$$

Then

$$\exp(x) + \exp(-x) \;\simeq\; 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots + 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \dots \simeq 2 + x^2$$

$$\exp(x) - \exp(-x) \;\simeq\; 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots - 1 + x - \frac{x^2}{2} + \frac{x^3}{6} + \dots \simeq 2x \qquad (9.59)$$

so that

$$\beta(x) \to \frac{2 + x^2}{2x} - \frac{1}{x} = \frac{1}{2}x$$

In both cases the function β is bounded and positive.

The expression above is also found at p. 617 in Logg, Mardal, and Wells [852]. Note that Hughes & Brooks [638] replace the costly 'coth' term by an asymptotic curve:
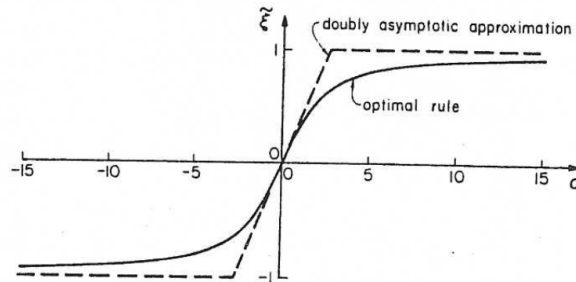
Taken from Hughes and Brooks [638] (1982). Here α is the Peclet number.

Note also that in Hughes and Brooks [638] (1982) the authors present the following formulation for the 2d case:

This approach will certainly prove to be rather expensive!

2. Codina (2000) (see Eq.39 of [288], see also Eq.2.64 of Donea and Huerta [363]) defines τ as follows:

$$\tau_2 = \left(\frac{2|\vec{v}|}{h} + \frac{4\kappa}{h^2} + \sigma\right)^{-1}$$

where σ is a reaction term (which we neglect in what follows). This equation can be re-written:

$$\tau_2 = \frac{h}{2|\vec{v}|}\left(1 + \frac{1}{\mathsf{Pe}}\right)^{-1}$$

3. In Donea and Huerta [363] we find the expression proposed by Shakib and co-workers (see Eq. 3.59 of [1230]):

$$\tau_3 = \left(\left(\frac{2|\vec{v}|}{h}\right)^2 + 9\left(\frac{4\kappa}{h^2}\right)^2 + \sigma^2\right)^{-1/2} = \frac{h}{2|\vec{v}|}\left(1 + \frac{9}{\mathsf{Pe}^2} + \left(\frac{h}{2|\vec{v}|}\sigma\right)^2\right)^{-1/2}$$

4. Braun [143] uses

$$\tau_4 = \frac{h}{|\vec{v}|\sqrt{15}} = \frac{h}{2|\vec{v}|}\frac{2}{\sqrt{15}}$$

This formula is independent of Pe and is also in Bochev, Gunzburger, and Shadid [111] (2004). It is also used in Hughes & Brooks (1982) [638] for pure advection problems in 1D and this value is attributed to Raymond & Garder (1976) [1116].

5. Following Tezduyar and Osawa [1336] (see also Appendix A of Thieulot (2011) [1343]):

$$\tau_5 = \left(\frac{2|\vec{v}|}{h} + \frac{1}{\theta\delta t} + \frac{\kappa}{h^2}\right)^{-1}$$

Crank-Nicolson: $\theta = 1/2$, CFL condition yields $\delta t = C\frac{h}{|\vec{v}|}$ so

$$\tau_5 = \left(\frac{2|\vec{v}|}{h} + \frac{2}{\delta t} + \frac{\kappa}{h^2}\right)^{-1} = \left(\frac{2|\vec{v}|}{h} + \frac{2v}{Ch} + \frac{\kappa}{h^2}\right)^{-1} = \frac{h}{2|\vec{v}|}\left(1 + \frac{1}{C} + \frac{1}{4\mathsf{Pe}}\right)^{-1}$$

Note that the velocity in the CFL criterion is the maximum velocity in the domain, while in all other expressions above for τ_i it is the velocity in the cell/element. By doing so, the relationship for τ_5 is actually only valid for the cell with the highest velocity. Also, this has been used in the context of SUPG methods for the Navier-Stokes equations, not advection-diffusion equations! This approach is then not considered further.

6. Franca, Hauke, and Masud [439] (2004) use the following formula which originates from Franca, Frey, and Hughes [438] (1992): if $0 \leq \mathsf{Pe} < 1$ then $\tau = \frac{h}{2|\vec{v}|}\mathsf{Pe}$ and if $\mathsf{Pe} \geq 1$ then $\tau = \frac{h}{2|\vec{v}|}$. Note however that their definition of the Peclet number includes a scalar parameter m which is the minimum between $1/3$ and and $2C_k$ where the calculation of this parameter is discussed in Remarks 4 and 5 of [438]. The authors conclude that for linear quadrilaterals the value $1/3$ should be used while for biquadratic elements $1/12$ should be preferred. The same approach is found in Brezzi, Bristeau, Franca, Mallet, and Rogé [159] (1992).

7. Knobloch [758] discusses other choices of τ.

Quoting Donea & Huerta: "It is obvious that τ must vanish when the mesh is refined (no stabilisation is necessary for a fine enough mesh)" and "Numerical experiments seem to indicate that for finite elements of order p the value of the stabilisation parameter should be approximately τ/p."

Let us define the dimensionless coefficient $\gamma_j = \frac{\tau_j}{h/2|\vec{v}|}$ and plot this quantity against the Peclet number:
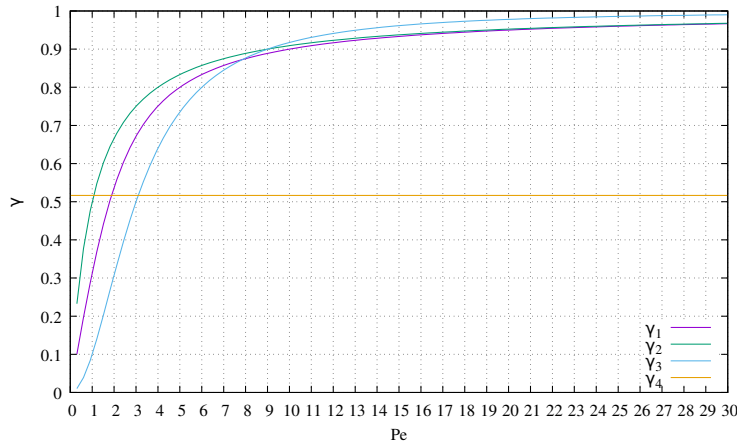
in images/supg/ folder. Note that gnuplot struggles with small values of Pe and the purple line should actually reach the origin.

In the case when the equation to be stabilised is a pure advection equation, then $\mathsf{Pe} \to \infty$ so

$$\gamma_1 \to 1/p \quad \gamma_2 \to 1 \quad \gamma_3 \to 1 \quad \gamma_4 \simeq 0.52 \quad \gamma_5 \to (1 + C^{-1})^{-1}$$

Note that in most versions above (except τ_1) τ does not depend on the order p of the basis functions. Is it an omission?

Finally it is worth mentioning that Donea and Huerta [363] argue that τ should ideally take different values at the corners and at the mid-side nodes for quadratic elements (p. 61).

9.6.3 Some remarks about Appendix A of Thieulot (2011) [1343]

In the DOUAR paper [144] or the FANTOM paper [1343], The advection matrix is simply modified and computed as follows:

$$(\mathbb{K}_a^e)_{SUPG} = \int_{x_k}^{x_{k+1}} (\vec{\mathcal{N}}^\star)^T \rho C_p \vec{v} \cdot \boldsymbol{B} dx \qquad \text{with} \qquad \vec{\mathcal{N}}^\star = \vec{\mathcal{N}} + \tau \vec{v} \cdot \boldsymbol{B}$$

Note that we can also write

$$(\mathbb{K}_a^e)_{SUPG} = \int_{x_k}^{x_{k+1}} \vec{\mathcal{N}}^T \rho C_p \vec{v} \cdot \boldsymbol{B} dx + \int_{x_k}^{x_{k+1}} \tau (\vec{v} \cdot \boldsymbol{B})^T \rho C_p (\vec{v} \cdot \boldsymbol{B}) dx$$

362

and we see that the SUPG method introduces and additional term that is akin to a diffusion term in the direction of the flow. This can be seen by looking at the advection matrix a regular grid of 1D elements of size h:

$$(\mathbb{K}_a^e)_{SUPG} = \mathbb{K}_a^e + \rho C_p \frac{\tau u^2}{h} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

The additional matrix has the same structure as the 1D diffusion matrix matrix in 9.1.

The parameter τ is chosen as follows [1336]:

$$\tau = \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + \frac{1}{\tau_3} \right)^{-1} \qquad \tau_1 = \frac{h}{2|\vec{v}|}, \quad \tau_2 = \theta \, \delta t, \quad \tau_3 = \frac{h^2 \rho C_p}{k} \tag{9.60}$$

where h is a measure of the element size and θ is related to the time discretisation scheme ($\theta = 1/2$ corresponds to a mid-point implicit scheme), and we can define $\gamma = \tau |\vec{v}|/h$ (see Appendix A of [1343]).

A typical test case for testing an advection scheme is the step advection benchmark (see for instance Donea & Huerta (2003) [363]). At $t = 0$, a field $T(x)$ is prescribed in a 1D domain of unit length. For $x \leq 1/4$ we have $T(x) = 1$ and $T(x) = 0$ everywhere else as shown on the following figure:
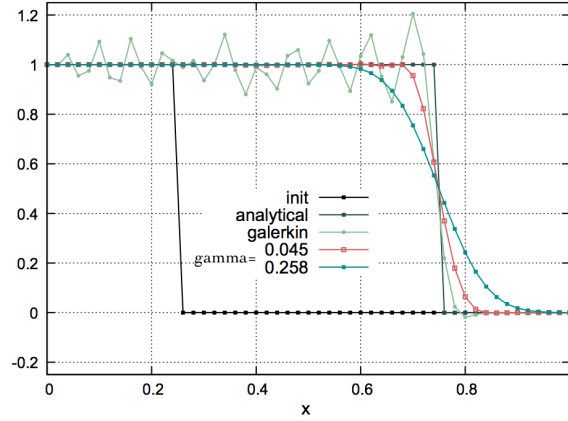
Taken and modified from Thieulot (2011) [1343]

The prescribed velocity is $v = 1$, 50 elements are used ($h = 0.02$) and 250 time steps are carried out with $\delta t = 0.1h/v = 0.002$ (CFL number of 0.1). Then it follows that $\tau_3 = \infty$ (no diffusion, i.e. $k = 0$) and

$$\tau = \left(\frac{2}{0.02} + \frac{2}{0.002} \right)^{-1} = \left(\frac{1}{0.01} + \frac{1}{0.001} \right)^{-1} = \left(\frac{1}{0.01} \left(1 + \frac{1}{10} \right) \right)^{-1} = 0.01 \left(\frac{11}{10} \right)^{-1} \simeq 0.009091$$

which yields $\gamma = 0.00909/0.02 = 0.4545...$ Using this value leads to a desired removal of the oscillations through a small amount of numerical diffusion. Braun [143] argues for a constant $\gamma = 1/\sqrt{15} = 0.258$ (citing Hughes & Brooks (1982) [638]), which effect is also shown in the figure above. This value is arguably too large and introduces an undesirable diffusion. Note that this same value is to be found in Bochev, Gunzburger, and Shadid [111] (2004). The authors then state that "for 1D pure advection problems, this choice maximizes the phase accuracy in the semidiscrete equation" and cite Raymond & Gardner (1976) [1116] as source.

Another classic example of advection testing is a 2d problem where (for example) a cylinder, a Gaussian and a cone are prescribed and advected with a velocity field (see for instance [363]).
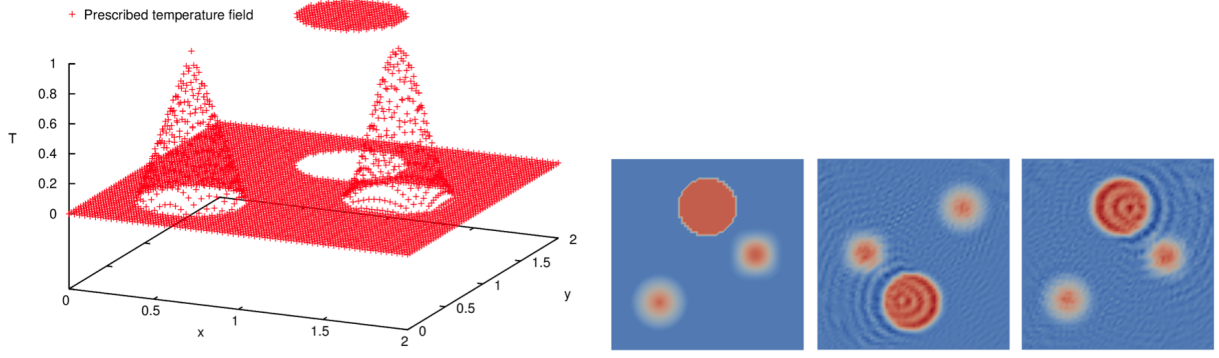
After a 2π rotation and in the absence of stabilisation we see that the temperature field showcases clearly visible ripples.

Remark. *Note that* ASPECT *originally did not rely on the SUPG formulation to stabilise the advection(-diffusion) equations[776]. It instead relied on the Entropy Viscosity formulation [531, 530]. It is only during the 6th Hackathon in May 2019 that the SUPG was introduced on the code. Note that the* ASPECT *implementation is based on the deal.II step 63[12].*

9.6.4 A note about linear elements - artificial diffusion

We have seen in Section 6.10 that the discretised advection-diffusion equation in 1D can be written in a manner akin to the Finite Difference method:

$$\frac{u}{2h}\left[\left(1 - \frac{1}{\mathsf{Pe}}\right)T_{i+1} + \frac{2}{\mathsf{Pe}}T_i - \left(1 + \frac{1}{\mathsf{Pe}}\right)T_{i-1}\right] = f$$

and we show in STONE 65 that the solution is far from accurate for $\mathsf{Pe} > 1$ (i.e. systems where advection dominates over diffusion). Let us ask ourselves the following question: could we come up with a modified version of the equation above which guarantees an exact solution on a uniform mesh of linear elements?

Essentially, we are looking for A, B and C such that not sure I understand this sentence

$$AT_{i+1} + BT_i + CT_{i-1} = f$$

One can show (see Donea & Huerta [363]) that a successful candidate formulation is[13]:

$$\frac{u}{2h}\left[(1 - \coth(\mathsf{Pe}))\,T_{i+1} + 2\coth(\mathsf{Pe})T_i - (1 + \coth(\mathsf{Pe}))\,T_{i-1}\right] = f$$

> I need to better understand this!

which can be arranged into the following form:

$$u\frac{T_{i+1} - T_{i-1}}{2h_x} - (\kappa + \tilde{\kappa})\frac{T_{i-1} - 2T_i + T_{i+1}}{h_x^2} = \frac{f}{\rho_0 C_p} \tag{9.61}$$

where $\tilde{\kappa}$ is known as artificial (numerical) diffusion (dissipation) given by

$$\tilde{\kappa} = \beta\frac{uh}{2} = \beta\kappa\mathsf{Pe} \qquad \text{with} \qquad \beta = \coth(\mathsf{Pe}) - \frac{1}{\mathsf{Pe}}$$

where Pe is the (dimensionless) Peclet number defined as:

$$\boxed{\mathsf{Pe} = \frac{uh}{2\kappa}}$$

---

[12]https://www.dealii.org/developer/doxygen/deal.II/step_63.html
[13]where we have used $\coth(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$

If $\mathsf{Pe} > 1$ we say the problem is advection-dominated, else if $\mathsf{Pe} < 1$ we say the problem is diffusion-dominated.

The $\beta(\mathsf{Pe})$ factor is shown in the following figure:
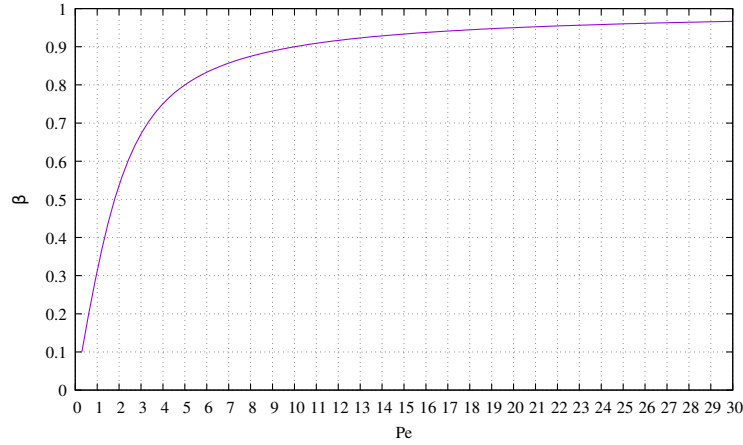
One can also define $\tilde{k} = \rho_0 C_p \tilde{\kappa}$ so that the discretised 1D advection-diffusion becomes:

$$\rho_0 C_p u \frac{T_{i+1} - T_{i-1}}{2h_x} - (k + \tilde{k}) \frac{T_{i-1} - 2T_i + T_{i+1}}{h_x^2} = f \tag{9.62}$$

or, in its continuous formulation:

$$\rho_0 C_p u \frac{dT}{dx} - (k + \tilde{k}) \frac{d^2 T}{dx^2} = f$$

$$\rho_0 C_p u \frac{dT}{dx} - \tilde{k} \frac{d^2 T}{dx^2} - k \frac{d^2 T}{dx^2} = f$$

$$\rho_0 C_p u \frac{dT}{dx} - \rho_0 C_p \beta \frac{uh}{2} \frac{d^2 T}{dx^2} - k \frac{d^2 T}{dx^2} = f \tag{9.63}$$

9.6.5 A note about linear elements - bubble functions & Petrov-Galerkin formulation

Let us assume that $u > 0$, i.e. advection goes from left to right. Node $i - 1$ is said to be on the upstream side of node i, and node $i + 1$ is on the downstream side of node . In Petrov FEM, instead of selecting the weight functions to be the same as the standard basis functions, we distort them as shown below.

The distortion is based on so-called bubble functions since they have zero values on the nodes and they are nonzero on elements' interiors. For instance one can take

$$\mathcal{N}_1^\star(r) = \frac{1}{2}(1 - r) - \frac{3}{4}\beta(1 - r^2) \tag{9.64}$$

$$\mathcal{N}_2^\star(r) = \frac{1}{2}(1 + r) + \frac{3}{4}\beta(1 - r^2) \tag{9.65}$$

as test/weight functions where β is a parameter that controls the amount of upwinding.

The elemental advection matrix is then given by:

365
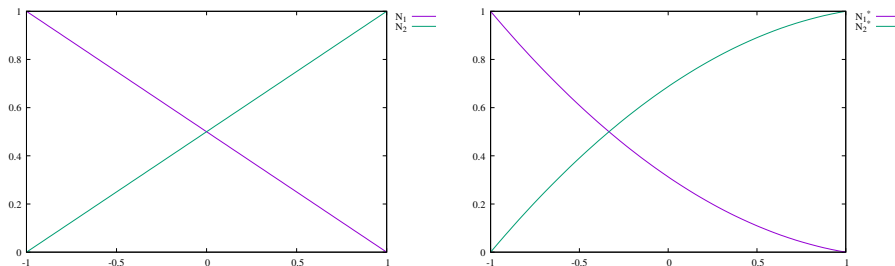
$$r = 2(x - x_k)/(x_{k+1} - x_k) - 1 = 2(x - x_k)/h - 1$$

then

$$dr = \frac{2}{h}dx$$

or

$$dx = \frac{h}{2}dr$$

$$
\begin{aligned}
\mathbb{K}_a^e &= \int_{x_k}^{x_{k+1}} (\vec{\mathcal{N}}^\star)^T \rho C_p \vec{v} \cdot \boldsymbol{B} \, dx \\
&= \frac{h}{2} \int_{-1}^{+1} (\vec{\mathcal{N}}^\star)^T \rho C_p \vec{v} \cdot \boldsymbol{B} \, dr \\
&= \frac{h}{2} \int_{-1}^{+1} \begin{pmatrix} \mathcal{N}_1^\star \\ \mathcal{N}_2^\star \end{pmatrix} \rho C_p u \begin{pmatrix} \dfrac{d\mathcal{N}_1}{dr} & \dfrac{d\mathcal{N}_2}{dr} \end{pmatrix} dr \\
&= \frac{h}{2} \rho C_p u \int_{-1}^{+1} \begin{pmatrix} \mathcal{N}_1 \\ \mathcal{N}_2 \end{pmatrix} \begin{pmatrix} \dfrac{d\mathcal{N}_1}{dr} & \dfrac{d\mathcal{N}_2}{dr} \end{pmatrix} dr + \frac{h}{2} \rho C_p u \int_{-1}^{+1} \begin{pmatrix} -\frac{3}{4}\beta(1-r^2) \\ +\frac{3}{4}\beta(1-r^2) \end{pmatrix} \begin{pmatrix} \dfrac{d\mathcal{N}_1}{dr} & \dfrac{d\mathcal{N}_2}{dr} \end{pmatrix} dr \\
&= \frac{h}{2} \rho C_p u \int_{-1}^{+1} \begin{pmatrix} \frac{1}{2}(1-r) \\ \frac{1}{2}(1+r) \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \end{pmatrix} dr + \frac{h}{2} \rho C_p u \int_{-1}^{+1} \begin{pmatrix} -\frac{3}{4}\beta(1-r^2) \\ +\frac{3}{4}\beta(1-r^2) \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \end{pmatrix} dr \\
&= \rho C_p u \frac{h}{8} \int_{-1}^{+1} \begin{pmatrix} -(1-r) & (1-r) \\ -(1+r) & (1+r) \end{pmatrix} dr + \frac{3h}{16} \beta \rho C_p u \int_{-1}^{+1} \begin{pmatrix} (1-r^2) & -(1-r^2) \\ -(1-r^2) & (1-r^2) \end{pmatrix} dr \\
&= \rho C_p u \frac{h}{8} \begin{pmatrix} -2 & 2 \\ -2 & 2 \end{pmatrix} + \frac{3h}{16} \beta \rho C_p u \begin{pmatrix} 4/3 & -4/3 \\ -4/3 & 4/3 \end{pmatrix} \\
&= \rho C_p u \frac{h}{4} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} + \frac{h}{4} \beta \rho C_p u \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{9.66}
\end{aligned}
$$

We see that this additional matrix is akin to $\mathbb{K}_d^e$ (albeit with a different diffusion coefficient).

Chapter 10

Solving the Stokes equations with the FEM

10.1 A quick tour of similar literature

- *Treatise on Geophysics*, Volume 7, Edited by D. Bercovici and G. Schubert: "Numerical Methods for Mantle Convection", by S.J. Zhong, D.A. Yuen, L.N. Moresi and M.G. Knepley. Note that it is a revision of the previous edition chapter by S.J. Zhong, D.A. Yuen and L.N. Moresi, Volume 7, pp. 227-252, 2007.

- *Computational Science I*, Lecture Notes for CAAM 519, M.G. Knepley, 2017. `https://cse.buffalo.edu/~knepley/classes/caam519/`

- *Numerical Modeling of Earth Systems - An introduction to computational methods with focus on solid Earth applications of continuum mechanics*, Th.W. Becker and B.J.P. Kaus, 2018. `http://www-udc.ig.utexas.edu/external/becker/Geodynamics557.pdf`

- *Myths and Methods in Modeling*, M. Spiegelman, 2000. `https://earth.usc.edu/~becker/teaching/557/reading/spiegelman_mmm.pdf`

In the case of an incompressible flow, we have seen that the continuity (mass conservation) equation takes the simple form $\vec{\nabla} \cdot \vec{v} = 0$. In other words flow takes place under the constraint that the divergence of its velocity field is exactly zero eveywhere (solenoidal constraint), i.e. it is divergence free.

We see that the pressure in the momentum equation is then a degree of freedom which is needed to satisfy the incompressibilty constraint (and it is not related to any constitutive equation) (see for example Donea and Huerta [363]). In other words the pressure is acting as a Lagrange multiplier of the incompressibility constraint.

Various approaches have been proposed in the literature to deal with the incompressibility constraint but we will only focus on the penalty method (section 10.7) and the so-called mixed finite element method 10.8.

10.2 The compatibility condition, LBB condition, or inf-sup condition

What follows is a rough attempt at making sense of it.

The Ladyžhenskaya-Babuška-Brezzi (LBB[1]) condition is a sufficient condition for a saddle point problem to have a unique solution. For saddle point problems coming from the Stokes equations, many discretizations (i.e. choices for the velocity and pressure polynomial spaces) are unstable, giving rise to artifacts such as spurious oscillations. The LBB condition gives criteria for when a discretization of a saddle point problem is stable. It also assures convergence at the optimal rate.

Bochev & Gunzburger [110] state:

> The terminology 'LBB' originates from the facts that this condition was first explicitly discussed in the finite element setting for saddle point problems by Brezzi[2] [158] and that it is a special case of the general weak-coercivity condition first discussed for finite element methods by Ivo Babuška[3] [40] and that, in the continuous setting of the Stokes equation, this condition was first proved to hold by Olga Ladyzhenskaya[4]; see [783].

Unfortunately, to quote Donea & Huerta [363]:

> In the finite element context, it is by no means easy to prove whether or not a given velocity-pressure pair satisfies the LBB compatibility condition.

Elman *et al.* state:

> [...] Choosing spaces for which the discrete inf-sup condition holds and is a delicate matter, and seemingly natural choices of velocity and pressure approximation do not work. [...] In general, care must be taken to make the velocity space rich enough compared to the pressure space.

By rich enough the authors essentially mean that the order of the polynomials used to represent velocity must be higher than the one used for pressure.

The LBB condition, or inf-sup condition can be proven in different ways, and standard techniques have been designed as listed in Boffi, Brezzi, and Fortin [114] (2008).

Elman *et al.* [394] state (p.129) that

> The inf-sup condition is a sufficient condition for the pressure to be unique up to constant in the case of an enclosed flow.

This can also be proven for other boundary conditions. This approach, based on the macro-element technique [1289] is explored in Appendix I.

It can be shown that, provided the kernel (null space) of matrix $\mathbb{G}$ is zero, the Stokes matrix is non-singular, that is $\vec{\mathcal{V}}$ and $\vec{\mathcal{P}}$ are uniquely defined, and the Schur complement matrix $\mathbb{S}$ is positive definite. Simply put, taking $\vec{\mathcal{V}} = \vec{0}$ in the discretised Stokes system without body forces yields $\mathbb{G} \cdot \vec{\mathcal{P}} = \vec{0}$ and implies that any pressure solution is only unique up to the null space of the matrix $\mathbb{G}$.

We know that the Schur complement matrix $\mathbb{S}$ is positive definite if and only if all of its eigenvalues are positive. One could then (numerically) compute the eigenvalues of $\mathbb{S}$ and check that these are indeed strictly positive to show that $\mathbb{S}$ is positive definite but that would prove very costly.

Another way is to see that $\mathbb{S}$ is positive definite only if $\ker(\mathbb{G}) = \{0\}$. Again to quote Donea & Huerta [363]: "If this is the case, the partitioned Stokes matrix is non-singular and delivers uniquely defined velocity and pressure fields. If this is not the case, a stable and convergent velocity field might be obtained, but the pressure field is likely to present spurious and oscillatory results." Note

---

[1]https://en.wikipedia.org/wiki/Ladyzhenskaya-Babuska-Brezzi_condition
[2]https://en.wikipedia.org/wiki/Franco_Brezzi
[3]https://en.wikipedia.org/wiki/Ivo_Babuska
[4]https://en.wikipedia.org/wiki/Olga_Ladyzhenskaya

that in the case of the $\boldsymbol{Q}_1 \times P_0$ element it has been shown that the multiple families of checkboard pressure modes actually lie in the kernel of $\mathbb{G}$. [1184, 1185]

We can look at this in a different manner, as explained in Elman, Silvester, and Wathen [394]: the unique solvability of the matrix system

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix} \tag{10.1}$$

is determined by looking at the homogeneous system

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} \tag{10.2}$$

or,

$$\begin{aligned} \mathbb{K} \cdot \vec{\mathcal{V}} + \mathbb{G} \cdot \vec{\mathcal{P}} &= \vec{0} \\ \mathbb{G}^T \cdot \vec{\mathcal{V}} &= \vec{0} \end{aligned} \tag{10.3}$$

To start, premultiply the first equation by $\vec{\mathcal{V}}^T$ and the second by $\vec{\mathcal{P}}^T$. The second yields $\vec{\mathcal{P}}^T \cdot \mathbb{G}^T \cdot \vec{\mathcal{V}} = (\vec{\mathcal{V}}^T \cdot \mathbb{G} \cdot \vec{\mathcal{P}})^T = \vec{0}$ which is present in the first equation so that it simplifies to $\vec{\mathcal{V}}^T \cdot \mathbb{K} \cdot \vec{\mathcal{V}} = \vec{0}$. Since $\mathbb{K}$ is positive definite, it follows that $\vec{\mathcal{V}} = \vec{0}$, implying unique solvability with respect to the velocity.

On the other hand, unique solvability with respect to the pressure is problematic. Substituting $\vec{\mathcal{V}} = \vec{0}$ in the system above gives $\mathbb{G} \cdot \vec{\mathcal{P}} = \vec{0}$, and implies that any pressure solution is only unique up to the nullspace of the matrix $\mathbb{G}$. The bottom line is that if Eq. (10.1) is to properly represent a continuous Stokes problem, then the mixed approximation spaces need to be chosen carefully. Specifically, we have to ensure that null($\mathbb{G}$) = {1} in the case of enclosed flow, and that null($\mathbb{G}$) = {0}, otherwise.

Gresho and Sani [517] state:

> LBB stable elements assure the existence of a unique solution to Stokes flow and assure convergence at optimal rate. [...] LBB-unstable elements may not converge, and if they do, they may not do so at the optimal rate.

10.3   A quick word about element families

The family of Taylor-Hood finite element spaces on triangular/tetrahedral grids is given by $\boldsymbol{P}_k \times P_{k-1}$ with $k \geq 2$, and on quadrilateral/hexahedral grids by $\boldsymbol{Q}_k \times Q_{k-1}$ with $k \geq 2$. This means that the pressure is then approximated by continuous functions.

These finite elements are very popular, in particular the pairs for $k = 2$, i.e. $\boldsymbol{Q}_2 \times Q_1$ and $\boldsymbol{P}_2 \times P_1$. The reason why $k \geq 2$ comes from the fact that the $\boldsymbol{Q}_1 \times Q_0$ (often referred to as $\boldsymbol{Q}_1 \times P_0$) and $\boldsymbol{P}_1 \times P_0$ are not stable elements (they are not inf-sup stable), as shown in John [688, p64] and [688, p67].

Remark. *Note that a similar element to $\boldsymbol{Q}_2 \times Q_1$ has been proposed and used succesfully used in Taylor and Hood [1326] (1973) and Hood and Taylor [622] (1974): it is denoted by $\boldsymbol{Q}_2^{(8)} \times Q_1$ since the center node ('$x^2 y^2$') and its associated degrees of freedom have been removed. It has also been proved to be LBB stable. These are also called Serendipity elements.*

On page 139 of Carey and Oden [224] (1986) we read:

> No triangle may have more than one side coincident with the boundary and no quadrilateral may have more than two sides coincident with the boundary.

10.4 Which velocity-pressure pair for Stokes? (quads & hexa)

10.4.1 The $Q_1 \times P_0$ pair

(tikz_q1p0.tex)
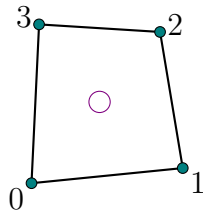
4 vel. nodes, 1 press. node   8 vel. nodes, 1 press. node
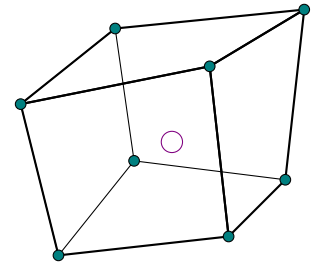
However simple it may look, the element is one of the hardest elements to analyze and many questions are still open about its properties. The element does not satisfy the inf-sup condition [637, p211]. In Gresho & Sani [517] it is labeled as follows: "slightly unstable but highly usable".

The $Q_1 \times P_0$ mixed approximation is the lowest order conforming approximation method defined on a rectangular grid. It also happens to be the most famous example of an unstable mixed approximation method [394, p235]. Boland and Nicolaides [118] (1984) and Boland and Nicolaides [119] (1985) show that it is not stable.

This element is discussed in Fortin (1981) [427], Fortin & Fortin (1985) [429] and in Pitkäranta & Saarinen (1985) [1063] in the context of multigrid use.

This element is plagued by so-called pressure checkerboard modes which have been thoroughly analysed, see for example Griffiths and Silvester [524] (1994), Chen, Pan, and Chang [238] (1995), Sani, Gresho, Lee, and Griffiths [1184] and Sani, Gresho, Lee, Griffiths, and Engelman [1185] (1981). These can be filtered out, see for example Chen, Pan, and Chang [238] (1995) or Lee, Gresho, and Sani [804] (1997), and explained in Section 16.8.

🛠 Relevant Literature: Fortin and Boivin [428] (1990), [520] (1985), Oden and Jacquotte [1007] (1984).

10.4.2 The $Q_2 \times Q_1$ pair

(tikz_q2q1.tex)
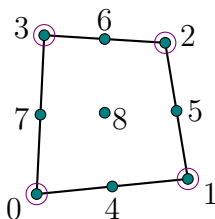
9 vel. nodes, 4 press. nodes   27 vel. nodes, 8 press. nodes
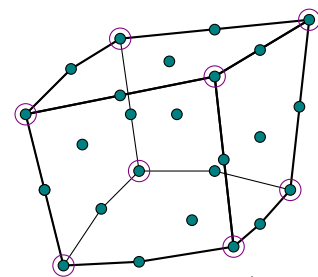
It belongs to the Taylor-Hood family of elements and satisfies the inf-sup (LBB) condition [637, p215]. Gresho & Sani [517, p554] write that in their opinion $div(\vec{v}) = 0$ is not strong enough. This

element, implemented in penalised form, is discussed in Bercovier & Engelman (1979) [84] and the follow-up paper [85].

It is the default of the Aspect code (see Appendix **??**). It is implemented in stone 18,21,48,91,120,...

10.4.3 The $Q_2 \times P_{-1}$ pair

all what followed has been temporarily copy pasted in stone 76 text

10.4.4 The $Q_k \times Q_{-(k-1)}$ pair

The $\boldsymbol{Q}_2 \times Q_{-1}$ element is shown in Table 3.13-2 of Gresho & Sani's book [517], and discussed in Section 3.13.6b of the book too. It is *not* LBB stable and has one chequerboard pressure mode.

It is used (alongside many other element pairs) in Christon, Gresho, and Sutton [274] (2002) in the context of a flow benchmark in a 2D box. The authors conclude that

> [...] the Q2-Q-1 element fared slightly better than the Q2-P-1 . Most surprising, though, were the good results obtained with the 'old' Taylor–Hood element, Q2-Q1 .

It is also used in Gresho and Sutton [518] (2002) on a similar benchmark setup (8:1 thermal cavity problem) along with $\boldsymbol{Q}_1 \times P_0$, $\boldsymbol{Q}_2 \times P_{-1}$ and $\boldsymbol{Q}_2 \times Q_1$. The authors state that Q2Q-1 has div-stability problems but

> produces excellent results and is still useful in general. [...] If the pesky-mode instability could be effiently dealt with, then the Q2xQ-1 element should be employed over the Q2xP-1 -especially in 3D (we believe).

Authors mention that it was also used in Vahl Davis and Jones [1390] (1983) and that it "performed extremely well."

In Zhang and Mu [1514] a special version of the $\boldsymbol{Q}_k \times Q_{-(k-1)}$ element is presented.

10.4.5 The $Q_1 \times P_0$-stab

Much has been written about the $Q_1 \times P_0$ element and the fact that it is not LBB-stable and that the pressure field contains a chequerboard mode that needs to be filtered out. It was the principal element used in computational geodynamics in codes such as Sopale, Citcom, Fantom, ... before being superseded by LBB-stable elements such as $Q_2 \times Q_1$ or $Q_2 \times P_{-1}$ [1345].

Many techniques have been proposed to stabilise this element but I here focus on those which keep the number of degrees of freedom unchanged, i.e. a matrix $\mathbb{C}$ is added to the Stokes matrix:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix}$$

More specifically I will focus on the pressure jump methods.

Note that in 3D the physical dimension of the $\mathbb{C}$ matrix is that of h^{dim}/η (i.e. $M^{-1}L^4T$) where h is the element size and η a viscosity. The Schur complement $\mathbb{S} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot G$ has obviously the same dimensions with $[\mathbb{G}] = L^2$ and $[\mathbb{K}] = MT^{-1}$.

As explained in Silvester & Kechkar [1245]: " The system [without the $\mathbb{C}$ matrix] is not strictly positive definite because of the zero coefficients on the diagonal. This fact makes pivoting necessary when solving [the system] by direct methods and limits the applicability of almost all iterative solution techniques. What is also well-known is that for certain combinations of the approximation velocity and pressure spaces, the uniqueness of the discrete solution may not be guaranteed. This is due to the occurrence of spurious pressure modes in the pressure approximation space."

The stability of mixed finite element methods boils down to properties of the null space of the matrix $\mathbb{G}$. An approximation is unstable if $\mathbb{G} \cdot \vec{\mathcal{P}} = \vec{0}$ where $\vec{\mathcal{P}}$ corresponds to some spurious pressure mode different from the constant value pressure. Note that if $\mathbb{G} \cdot \vec{\mathcal{P}} = \vec{0}$, then $(\vec{0}, \vec{\mathcal{P}})^T$ is a null vector of the homogeneous system.

The basic idea behind stabilization is to relax the incompressibility constraint in a special way so that this vector is no longer a null vector of the resulting coefficient matrix, and the discrete solutions satisfy rigorous error bounds. In other words the idea consists in regularising the system by replacing the zero block by an appropriate positive semi-definite matrix $-\mathbb{C}$ [1245].

We will here look at so-called local and global jump methods (and their various flavours), the macro-element method, as well as the penalty method (which is not really a stabilisation method, as we will see).

- global jump stabilisation: Hughes, Franca, and Balestra [640] (1987), Norburn and Silvester [1002] (1998), Douglas and Wang [366] (1989), Christon and Cook [273] (2001), Cao [223] (2003), Eguchi [385] (2003), Y. Cao [1477] (2006)

- local jump stabilisation: Silvester and Kechkar [1245] (1990), Kechkar and Silvester [727] (1992), Vincent and Boyer [1425] (1992), Cao [223] (2003), Qin and Zhang [1091] (2007), Christon [272] (2002), Christon and Cook [273] (2001), Liao and Silvester [833] (2013), Y. Cao [1477] (2006)

- stabilisation through macro-elements: Fortin and Boivin [428] (1990), LeTallec and Ruas [820] (1986), LeTallec [819] (1981)

In effect, these jump stabilisation techniques provide an a-priori filter for the weakly unstable pressure modes associated with the $Q_1 \times P_0$ element.

Consistency: in Barth, Bochev, Gunzburger, and Shadid [54]:"we should define what we mean by a **consistent method**; perhaps a more apt terminology would be variationally consistent. In standard usage, consistency of numerical schemes for partial differential equations requires that the pointwise truncation error vanish as the grid size goes to zero; i.e., if one substitutes a smooth solution of the partial differential equation into the numerical scheme, then the residual is at least o(h), where h denotes the grid size. Finite element schemes are not, in general, consistent in this sense. However, for standard finite element methods, sufficiently smooth exact solutions of the partial differential equations exactly satisfy the variational equation that defines the discrete finite element equations. This is what we mean by a consistent finite element scheme. This allows us to differentiate between the methods we consider in this paper and methods which are not consistent in this latter sense. For example, penalty methods for the Stokes problem are not consistent finite element methods since substitution of an exact solution into the discrete equations leaves a residual that is proportional to the penalty parameter. Thus, we consider only methods that do not suffer from this type of variational inconsistency."

As explained in Elman book: "to ensure consistency we require $\vec{1} \in null(\mathbb{C})$ (this precludes the use of inconsistent 'penalty methods') and we require $\vec{p}^T \cdot \mathbb{C} \cdot \vec{p} > 0$ for all spurious pressure modes $\vec{p} \neq \vec{1}$ in $null(\mathbb{G})$."

define jump operator !! looking at sike90, I realise that I don't understand how the(ir) jump operator works in practice. Eq on page 78?

The driving question behind all this, besides my wanting to understand these stabilisation schemes better, is the fact that 1) none of the existing publishing literature seems to address the problem of large and/or sharp viscosity contrasts/variations; 2) almost all papers deal with regular meshes and rectangular elements.

Penalty The conventional way of computing a regularisation matrix $\mathbb{C}$ is to use a penalty formulation. In the framework presented in Silvester and Kechkar [1245] (1990), the standard penalty method corresponds to the specific choice of

$$C(q^h, p^h) = \epsilon \int_\Omega q^h p^h \, dV = \epsilon \mathbb{M}_p \tag{10.4}$$

with $\epsilon > 0$ and $\mathbb{M}_p$ is the pressure mass matrix. For a regular grid of squares with size h, it follows that

$$C_{ij} = 0 \quad \text{if } i \neq j \qquad \text{and} \qquad C_{ii} = \epsilon \int_{\Omega_i} dV = h^2 \epsilon$$

so that the stabilisation matrix is diagonal:

19	20	21	22	23	24
13	14	15	16	17	18
7	8	9	10	11	12
1	2	3	4	5	6
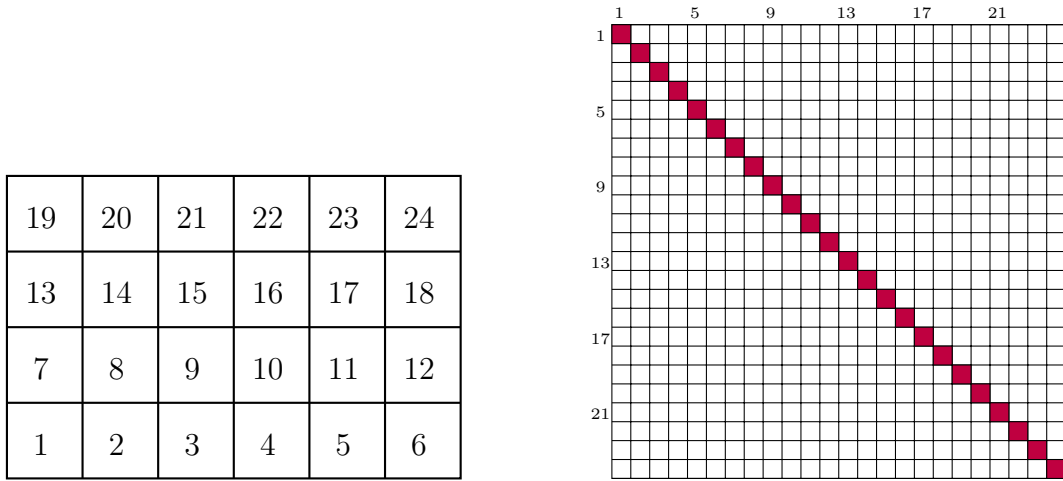
Two-dimensional grid composed of 6 × 4 elements on the left and the resulting sparsity pattern of the $\mathbb{C}$ matrix on the right.

In Zienkiewicz & Taylor (Section 4.8.2) the matrix is also written as $\mathbb{C} = \epsilon h^2 \mathbf{1}$ where $\mathbf{1}$ is the unit matrix.

It is stressed here that the penalty technique does not stabilise an unstable mixed method [1245]. A small penalty parameter means that the original problem is solved quite accurately.

See Cuvelier, Segal, and Steenhoven [317] (1986) for some more details about the penalty method. The approach above is similar to the one presented in Section **??**. The only difference is that instead of replacing the pressure in the momentum equation by $p = \lambda \vec{\nabla} \cdot \vec{v}$ we keep both velocity and pressure as unknowns and we take $\epsilon = \lambda^{-1}$. Since normally $\lambda >> \eta$ then $\epsilon = \lambda^{-1}$ must be small. As explained in Silvester & Kechkar [1245], "despite its theoretical attraction, the penalty technique breaks down in practice because of its sensitivity to the particular choice of penalty parameter".

The Stokes system for a single element then writes

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\epsilon \mathbb{M}_p \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix}$$

The second line yields

$$\mathbb{G}^T \cdot \vec{\mathcal{V}} - \epsilon \mathbb{M}_p \cdot \vec{\mathcal{P}} = \vec{h}$$

or,

$$\vec{\mathcal{P}} = \frac{1}{\epsilon} \mathbb{M}_p^{-1} \cdot (\mathbb{G}^T \cdot \vec{\mathcal{V}} - \vec{h})$$

which can be re-introduced in the first line:

$$\mathbb{K} \cdot \vec{\mathcal{V}} + \mathbb{G} \cdot \frac{1}{\epsilon} \mathbb{M}_p^{-1} \cdot (\mathbb{G}^T \cdot \vec{\mathcal{V}} - \vec{h}) = \vec{f}$$

or,

$$\left(\mathbb{K} + \frac{1}{\epsilon} \mathbb{G} \cdot \mathbb{M}_p^{-1} \cdot \mathbb{G}^T \right) \cdot \vec{\mathcal{V}} = \vec{f} + \frac{1}{\epsilon} \mathbb{G} \cdot \mathbb{M}_p^{-1} \cdot \vec{h}$$

This elimination could be carried out element by element so that one only solves for the velocity degrees of freedom.

    - reduced integration? sike90 does not say anything about it.

    - condition number explodes since $\frac{1}{\epsilon} >> \eta$

Dohrmann and Bochev [358] (2004) state: "Penalty methods are another category of non-residual based regularizations. They, however, differ from stabilized methods in the sense that application of a penalty does not circumvent the inf–sup condition and only serves to uncouple pressure from velocity. In this sense, penalty methods should be viewed as solution, rather then stabilization procedures for the mixed equations."

Global jump This method is explained in Silvester & Kechkar (1990) and the authors state that it was introduced by Hughes & Franca [640] in which a general theoretical framework for analysing global stabilisation techniques is presented. Using this framework, optimum rates of convergence for the $Q_1 \times P_0$ method stabilised with global jumps are established.

The global jump stabilisation formulation introduces a pressure diffusion operator that perturbs the incompressibility constraint. The global jump formulation insures mass conservation in a global sense since the null space of the stabilising matrix constrains the constant-pressure vector. However, the global jump stabilisation smears the div-free constraint over a small region, i.e., the divergence is not zero at the element level [275].

Consider the stabilisation term

$$C(q^h, p^h) = \beta h \sum_{s=1}^{N_s} \int_{\partial \Omega_s} [\![q^h]\!] [\![p^h]\!] ds \tag{10.5}$$

in which h is the mesh parameter (defined locally), $[\![.]\!]$ is the jump operator, and $\beta > 0$ is a stabilising parameter. The summation is over *all* interior inter-element edges.

To illustrate, consider element 9 in the mesh consisting of equally sized squares represented here:

(tikz_globaljump.tex)

19	20	21	22	23	24
13	14	15	16	17	18
7	8	9	10	11	12
1	2	3	4	5	6

Element 9 has four direct neighbours: 3, 8, 10, and 15. The stabilisation term for this element involves the sum over its four neighbours:

$$\beta h \sum_{s=1}^{4} \int_{\partial \Omega_s} [[q^h]][[p^h]]ds = \beta h^2 [(p_9 - p_3) + (p_9 - p_8) + (p_9 - p_{10}) + (p_9 - p_{15})]$$

$$= \beta h^2 (4p_9 - p_3 - p_8 - p_{10} - p_{15})$$

The integral along each edge is simply the pressure difference across the edge multiplied by the edge surface/length which happens to be constant in this case. This means that the in the matrix $\mathbb{C}$, there will be entries on the 9th line at columns 3, 8, 10, and 15.

Be careful, let us now turn to element 6: it has 2 neighbours (5 and 12), so that the stabilisation term for this element involves the sum over its two neighbours:

$$\beta h^2 [(p_6 - p_5) + (p_6 - p_{12})] = \beta h^2 (2p_6 - p_5 - p_{12})$$

And looking now at element 23: it has three neighbours (17, 22, and 24), so that the stabilisation term for this element involves the sum over its three neighbours:

$$\beta h^2 [(p_{23} - p_{17}) + (p_{23} - p_{22}) + (p_{23} - p_{24})] = \beta h^2 (3p_{23} - p_{17} - p_{22} - p_{24})$$

The resulting assembled $\mathbb{C}$ matrix is shown here:

19	20	21	22	23	24
13	14	15	16	17	18
7	8	9	10	11	12
1	2	3	4	5	6
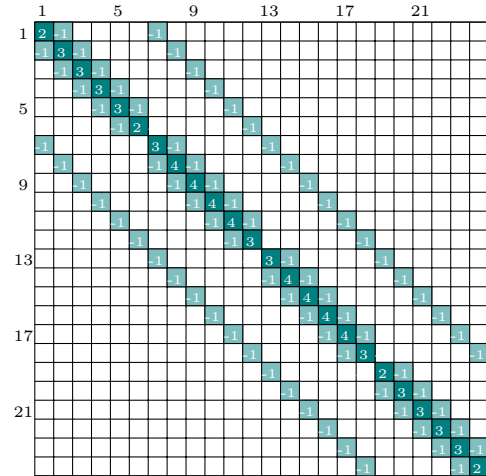
As stated in [1245], "for a natural numbering strategy the stabilisation matrix C is pentadiagonal".

The global jump stabilisation is effective in practice, although a careful choice of the parameter β is required in order to prevent a loss of accuracy in the solution. According to Silvester and Kechkar [1245] the only other deficiency is the fact that the global nature of the jump terms makes the method awkward to implement into existing codes [5].

Let us now consider the following chequerboard pressure mode:

+1	-1	+1	-1	+1	-1
-1	+1	-1	+1	-1	+1
+1	-1	+1	-1	+1	-1
-1	+1	-1	+1	-1	+1

---

[5] I don't understand this remark

We find

$$\beta h^2(4p_9 - p_3 - p_8 - p_{10} - p_{15}) = \beta h^2(4\cdot(+1) - (-1) - (-1) - (-1) - (-1)) = 8\beta h^2 \neq 0$$

Remark. *In Cao [223] (2003) the parameter β is set to 1.*

Remark. *Note that this approach is somewhat linked to the idea of a pressure smoother via a discrete Laplace[6]: The Discrete Laplace operator is often used in image processing e.g. in edge detection and motion estimation applications. The discrete Laplacian is defined as the sum of the second derivatives Laplace and calculated as sum of differences over the nearest neighbours of the central pixel. Here is an example of a 2D filter and it is of course similar in nature to the Global stabilisation stensil matrix:*

$$D^{(1)} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

Note that other filters can also be found although we will not use them here in this context:

$$D^{(2)} = \begin{bmatrix} -0.5 & -1 & -0.5 \\ -1 & +6 & -1 \\ -0.5 & -1 & -0.5 \end{bmatrix} \qquad D^{(3)} = \begin{bmatrix} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \qquad D^{(4)} = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

Eguchi (2003) adds a linear form to the global stab $\mathbb{C}$ to suppress the pressure nullspace.

Local jump According to Silvester and Kechkar [1245], the deficiencies of the global jump method can be overcome by a straightforward modification. Assume that the elements in can now be assembled into N_m disjoint macro-elements of 2×2 elements, as shown in gray on the following figure:

Consider now the bilinear form given by

$$C(q^h, p^h) = \beta h \sum_{m=1}^{N_m} \sum_{i=1}^{4} \int_{\partial\Omega_m} [\![q^h]\!][\![p^h]\!] ds$$

where the first summation is over all 2×2 macroelements, and the second summation runs over all inter element edges strictly within each macrolelement.

The form of the stabilisation matrix $\mathbb{C}$ is similar to that above except that there is now a local basis.

For instance, considering again element 9, it now belongs to the second macro-element and therefore only 'sees' neighbours 3 and 10. The resulting $\mathbb{C}$ matrix is shown on the figure here after and its structure is obviously different than in the global stabilisation case, albeit also pentadiagonal.

---

[6] http://en.wikipedia.org/wiki/Discrete_Laplace_operator
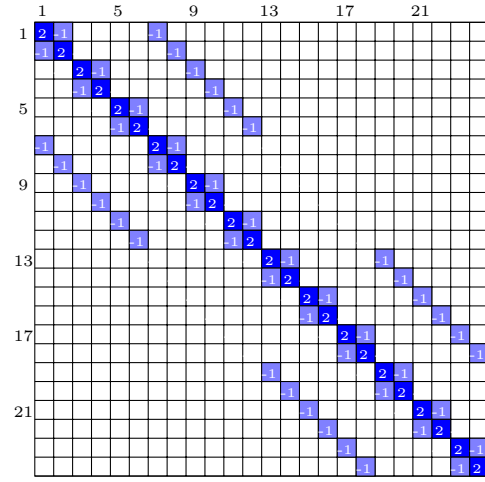
Remark. *Obviously, one could re-number the elements so that matrix* $\mathbb{C}$ *is block diagonal: macroelement 1 would contain elements 1,2,3,4, macroelement 2 would contain elements 5,6,7,8, etc ...*

According to Silvester and Kechkar [1245] or Chibani and Kechkar [251], the advantages of this local method over the global jump formulation are:

1. implementation is more straightforward because for assembly purposes each 2×2 block of elements can be treated as a single macroelement [7]

2. mass is conserved locally (over a macroelement), using the global jump formulation mass is only conserved globally

3. robustness is improved in the sense that the discrete velocity solution is less sensitive to the magnitude of β, the influence of the stabilisation matrix being localised (will need to be tested numerically!).

Remark. *The globally stabilised formulation corresponds to the extreme case of a local stabilisation based on a single macro-element [517].*

One of the features of the local stabilisation is that if the discrete incompressibility constraints are added together then the jump terms sum to zero in each macro element. Indeed, let us consider the following macro element:

The corresponding matrix (making abstraction of the β term) writes:

$$
h^2 \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}
$$

and the row/column sum of its entries is always null. Also, $\mathbb{C}$ is obviously positive semi-definite [1245].

---

[7] same here, not sure what they mean by this

Gresho & Sani [517] state:" This is crucially important to the success of the method since it implies that the local incompressibility of the $Q_1 \times P_0$ method is retained after stabilisation (albeit over macro-elements). It also suggests that a good strategy when constructing the partition is to form macro-elements containing as few elements as possible. Once a suitable macro-element partitioning has been formed, the local stabilisation matrices can be calculated by running through the component elements, summing jump contributions corresponding to the internal edges."

If one now considers the following irregular macro-element,
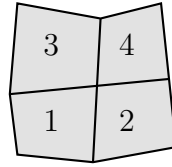
the corresponding matrix is given by[8]

$$
\tilde{h}
\begin{pmatrix}
h_{12} + h_{13} & -h_{12} & -h_{13} & 0 \\
-h_{12} & h_{12} + h_{24} & 0 & -h_{24} \\
-h_{13} & 0 & h_{13} + h_{34} & -h_{34} \\
0 & -h_{24} & -h_{34} & h_{24} + h_{34}
\end{pmatrix}
\begin{pmatrix}
p_1 \\
p_2 \\
p_3 \\
p_4
\end{pmatrix}
$$

where h_{ij} is the length/surface of the edge betweens elements i and j. The reference length $\tilde{h}$ may be computed by simply defining it to be the average diameter of the constituent elements.

Remark. *In three dimensions, the $2 \times 2 \times 2$ block is the obvious starting point for stabilising $Q_1 \times P_0$ [517].*

Perhaps the most serious potential drawback of the local framework is that stability is only guaranteed if the stabilisation parameter β is bigger than some critical value β_0, which needs to be estimated.

It can be estimated that $\beta = 1/4$ in 2D and $\beta = 1/6$ in 3D (see Gresho & Sani [517, p636] for a detailed derivation, see also Vincent and Boyer [1425] (1992)).

Silvester & Kechkar state: " The advantages of the stabilisation procedures over the penalty method are especially relevant to the discretisation of 3D incompressible flow problems, since iterative solution methods have to be used. Similar stabilisation techniques to those described here are applicable to the three-dimensional version of the $Q_1 \times P_0$ mixed method ".

The conclusion in [1245] is as follows: The local jump formulation proves to be an efficient method for a priori filtering of spurious pressure modes. It cleanly stabilises the $Q_1 \times P_0$ mixed method without compromising its simplicity and resulting efficiency; in particular, it is very robust with respect to the magnitude of the stabilisation parameter.

It is reported in Gresho & Sani [517] that when using an iterative solver, iteration counts are only independent of the grid in the stabilised cases: using the raw $Q_1 \times P_0$ method the iteration counts significantly increase with decreasing h. Note that the deterioration of the condition number of the matrix with decreasing h is worse in 3D than in 2D (but bear in mind that one almost always use higher resolutions in 2D than in 3D, so it does not help). 3D is also discussed in [**chsu97**].

A way to look at the global vs. local stabilisation schemes is presented on the following figure from Christon (2002) [272]:

---

[8]I suspect it should involve the normal vectors to the edges ...?
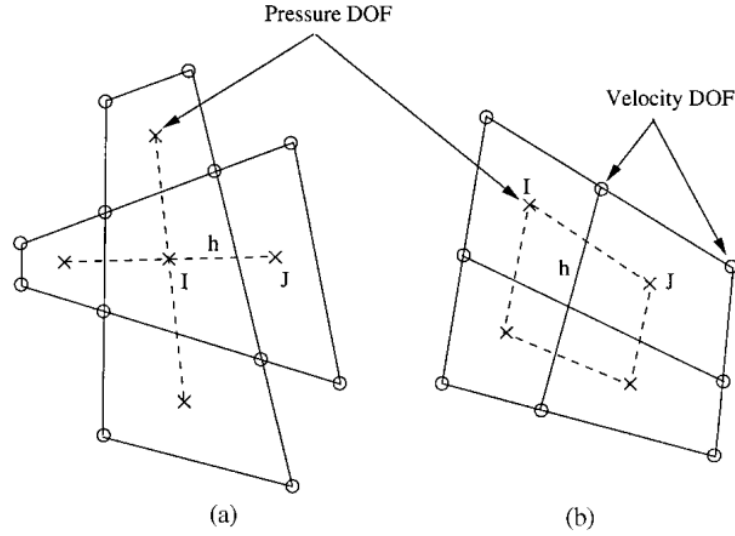
Element configuration for pressure stabilization: (a) global jump; (b) local jump.

Remark. *The locally stabilised Q1P0 and P1P0 elements have been analysed in Kechkar & Silvester [727]. Penalty, global and local approaches are mentioned in Vincent & Boyer (1992) but only the local jump stabilisation is used.*

Silvester (1994) investigates the value for β for Q1P0 and arrives at $0.0615 \leq \beta \leq 0.25$. Chang & Sugiyama (1997) report that "a value between 0.01 and 0.1 appears to work well for most applications conducted so far". In [394] the authors state that $\beta = \frac{1}{4}$ is an idea value which "ensures stability independently of the rectangle aspect ratio".

Norburn & Silvester (1998) state: "Although the 'optimal parameter' (the value of β which minimises the discretisation error on a given mesh) is impossible to determine a priori, good parameter choices (usually over-estimating the optimal parameter) can be found by minimising the condition number of the pressure Schur complement matrix. [...] The motivation for choosing the parameter value β which minimises the condition number (ratio of largest to smallest eigenvalue) of the pressure Schur complement is that this quantity roughly determines the rate of convergence of Uzawa type iteration methods." They conclude that $\beta \simeq 0.1$ is appropriate for P1P0.

Rather interestingly we see that choosing the 'best' β is primarily based in the literature on the Schur complement condition number, and less on the accuracy of the solutions (although it is sometimes assessed, see Fig. 4 of Norburn and Silvester [1002] (1998)).

Finally it is worth mentioning a recent paper by Chibani and Kechkar [251] (2020) who present modified local jump stabilisation schemes which effectively only take 2 or 1 one pressure jump into account instead of 4 per macro-element.

In three dimensions the matrix $\mathbb{S}$ is obtained by assembling the submatrices for the macroelements which, in general, are made up of 8 (i.e., $2 \times 2 \times 2$) adjacent elements, as shown in the following sketch [**chsu97**]:
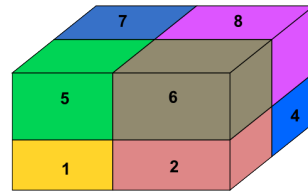
Numbering of hexahedrons in a macroelement.

Element No. 3 is behind No. 1 and below No. 7.

For a macroelement of 8 elements as ordered in the above sketch, the submatrix is defined below:

379

$$
\mathbb{S} = \begin{pmatrix}
h_{12}+h_{13}+h_{h15} & -h_{12} & -h_{13} & 0 & -h_{15} & 0 & 0 & 0 \\
-h_{21} & h_{21}+h_{24}+h_{26} & 0 & -h_{24} & 0 & -h_{26} & 0 & 0 \\
-h_{31} & \ldots & & & & & & \\
0 & \ldots & & & & & & \\
-h_{51} & \ldots & & & & & & \\
0 & \ldots & & & & & & \\
0 & \ldots & & & & & & \\
0 & \ldots & & & & & &
\end{pmatrix}
$$

in which h_{ij} is the length scale for elements 'i' and 'j' in the macroelement. The matrix $\mathbb{S}$ is symmetric, since $h_{ij} = h_{ji}$.

Two approaches are possible for computing the length scale h_{ij}:

- the square root of the interior inter-element surface area between elements 'i' and 'j'

- the quotient of the interior inter-element surface area divided by the cube root of the average element volume of the macroelement.

Macro-element   One source for this stabilisation approach is Section 5.3.2 of the book by Elman, Silvester and Wathen [394]. The $\mathbb{C}$ matrix for the 2D macroelement is shown to be:

$$
\mathbb{C} = \beta \begin{pmatrix}
1 & -1 & 1 & -1 \\
-1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 \\
-1 & 1 & -1 & 1
\end{pmatrix}
$$

and the authors suggest $\beta = \frac{1}{4} h_x h_y$.

Also see [428] (1990) .

Numerical scaling of $\mathbb{C}$ Since the matrix $\mathbb{K}$ contains the viscosity, it is to be expected that the magnitude of the entries in matrix $\mathbb{C}$ must somehow follows the values in the $\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G}$ term. Indeed the Schur complement is $\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}$ so if the entries in $\mathbb{C}$ are wrongly scaled it will either have no effect at all or will alter the solution too much. This is indeed what is advocated in Christon (2002) [272]:

> The global jump stabilization formulation attempts to control the jump in pressure across element boundaries, and results in a PPE that is perturbed by a 'pressure-diffusion' operator. The off-diagonal entries in the global jump stabilization matrix are defined as
>
> $$ S_{IJ} = \beta \frac{|[C^T M_L^{-1} C]_{IJ}|}{\Gamma_{IJ}} \int_{\Gamma_{IJ}} [\![\psi_I]\!][\![\psi_J]\!]\, d\Gamma, \qquad (23) $$
>
> where I and J identify adjacent elements that share a common face as shown in Figure 1(a). Here, Γ_{IJ} represents the shared inter-element boundary, $[\cdot]$ is the jump operator, and β is a non-dimensional scaling parameter. For the $Q1Q0$ element, the pressure approximation is piecewise constant with $\psi_I = 1$ inside Ω_e and zero outside. In two dimensions, Γ_{IJ} represents the length of the element edge shared by element I and J, and in three dimensions, it represents the area of

In the paper S stands for $\mathbb{C}$, and C stands for $\mathbb{G}$. Also, because Christon is solving the N-S equations, and because of his algorithmic choice to do so, there is a velocity mass matrix where our $\mathbb{K}$ block resides. As such his implementation showcases the lumped mass matrix in the equation above rather than the lumped $\mathbb{K}$ matrix.

In the paper Christon states that the PPE term $C^T M_L^{-1} C$ is symmetric. This is obviously true since it is a scalar for the $Q_1 \times P_0$ element. Am I missing something here?

Using our notations, the off-diagonal entries of the stabilisation matrix $\mathbb{C}$ for the global approach becomes:

$$
\mathbb{C}_{ef} = -\beta (\mathbb{G}^T \cdot \mathbb{K}_L^{-1} \cdot \mathbb{G})_e \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma \qquad \text{for } e \neq f \qquad (10.6)
$$

where e and f identify adjacent elements that share a common face, Γ_{ef} represents the shared inter-element boundary (a length in 2D, a surface in 3D), β is a non-dimensional scaling parameter, and $\mathbb{K}_L$ is the row-wise lumped $\mathbb{K}$ matrix[9]. For the $Q_1 \times P_0$ element, the pressure approximation is piecewise constant with $\psi_i = 1$ inside the element and zero outside.

The inclusion of the $\mathbb{G}^T \cdot \mathbb{K}_L^{-1} \cdot \mathbb{G}$ term in the stabilisation yields proper dimensionality of the stabilization matrix (the integrand is dimensionless so that the dimensions of $\mathbb{C}$ are those of the elemental Schur complement block), accounts for scaling due to irregular elements, and still preserves the symmetry [272].

Finally, the diagonal element of C for element e is computed as

$$\mathbb{C}_{ee} = \beta(\mathbb{G}^T \cdot \mathbb{K}_L^{-1} \cdot \mathbb{G})_e \sum_{f \neq e} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma$$

We find that the sum of the terms on the row corresponding to e is indeed zero, consistency is ensured even for non-rectangular elements.

There is however a major problem with this approach: even when the viscosity is constant in the domain, Eq. (10.6) does not yield a symmetric matrix if elements are not identical in shape. Since $\mathbb{C}_{ef} \propto (\mathbb{G}^T \cdot \mathbb{K}_L^{-1} \cdot \mathbb{G})_e$ then $\mathbb{C}_{ef} \neq \mathbb{C}_{fe}$! I then suspect that Christon's notation $|C^T M_L^{-1} C|_{IJ}$ indicates that some care must be taken so as to ensure $S_{IJ} = S_{JI}$ but it is not further specified. We will then have to figure this out.

Let us consider the case of square elements of size $h_x = h_y = h$. Then the $\mathbb{K}_e$ and $\mathbb{G}_e$ matrices are given by:

$$\mathbb{K}_e = \begin{pmatrix} 1 & 0.25 & -0.5 & -0.25 & -0.5 & -0.25 & 0 & 0.25 \\ 0.25 & 1 & 0.25 & 0 & -0.25 & -0.5 & -0.25 & -0.5 \\ -0.5 & 0.25 & 1 & -0.25 & 0 & -0.25 & -0.5 & 0.25 \\ -0.25 & 0 & -0.25 & 1 & 0.25 & -0.5 & 0.25 & -0.5 \\ -0.5 & -0.25 & 0 & 0.25 & 1 & 0.25 & -0.5 & -0.25 \\ -0.25 & -0.5 & -0.25 & -0.5 & 0.25 & 1 & 0.25 & 0 \\ 0 & -0.25 & -0.5 & 0.25 & -0.5 & 0.25 & 1 & -0.25 \\ 0.25 & -0.5 & 0.25 & -0.5 & -0.25 & 0 & -0.25 & 1 \end{pmatrix} \qquad \mathbb{G}_e = h \begin{pmatrix} +1/2 \\ +1/2 \\ -1/2 \\ +1/2 \\ -1/2 \\ -1/2 \\ +1/2 \\ -1/2 \end{pmatrix}$$

so that the Schur complement is

$$\mathbb{S}_e = \mathbb{G}_e^T \cdot \tilde{\mathbb{K}}_e^{-1} \cdot \mathbb{G}_e = \frac{2}{3} h^2$$

In that case we almost recover the expression of for example the macro-element.

Dealing with viscosity contrasts/large variations This topic is almost never discussed as many papers consider the standard Stokes equations with $\eta = 1$ (and also regular meshes made of identical elements). This is not problematic in engineering where often the fluid in question has a constant viscosity (or the equations have been rendered dimensionless). However, in geodynamical applications we know that the viscosity field can showcase very sharp gradients (sinking/rising objects, shear bands, free surface, ...). In what follows we assume that each element e has an effective viscosity η_e.

The scaling of the $\mathbb{C}$ matrix in the previous section is not formulated when viscosity contrasts from one element to the other are present. For example scaling the row entries of the $\mathbb{C}$ matrix by

---

[9]The quantity $\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G}$ for $Q_1 \times P_0$ elements is a scalar, which is rather convenient as it gives in a simple way the scaling for the stabilisation term. However the inverse of $\mathbb{K}$ is costly so there is a cheaper alternative which consists in lumping it so that it becomes diagonal and its inverse is then trivial.

the element viscosity still yields a structurally symmetric matrix, but not a numerically symmetric one which is problematic since we have seen that $\mathbb{C}$ must be semi-positive definite. Some form of viscosity averaging must then take place between adjacent elements so that the contribution from element e to f is exactly the same as f to e

In order for the stabilisation to remain consistent it must satisfy $\mathbb{C} \cdot \vec{1} = 0$, i.e. it should have zero effect on a constant pressure field, which then forces the sum of the entries for each row (or column) to be null. This requirement makes the above viscosity averaging idea very difficult in practice in the global case (satisfying both symmetry and consistency).

Let us start by defining the elemental Schur complement

$$\tilde{\mathbb{S}}_e = |(\mathbb{G}^T \cdot \mathbb{K}_L^{-1} \cdot \mathbb{G})_e|$$

and then

$$\tilde{\mathbb{S}}_{ef} = \phi(\tilde{\mathbb{S}}_e, \tilde{\mathbb{S}}_f)$$

where ϕ is a function to be specified later such that $\phi(x, y) = \phi(y, x)$. The local and global jump stabilisation can then be formulated as follows:

$$\mathbb{C}_{ef} = -\beta \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma \qquad \text{for } e \neq f \tag{10.7}$$

supplemented by

$$\mathbb{C}_{ee} = \beta \sum_{f \neq e} \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma$$

In this case the matrix $\mathbb{C}$ is symmetric and consistent!

Recap : For each element compute its (scalar) Schur complement

$$\tilde{\mathbb{S}}_e = |(\mathbb{G}_e^T \cdot \tilde{\mathbb{K}}_e^{-1} \cdot \mathbb{G})|$$

We here assume that elements can have different viscosities and/or shape so that $\mathbb{S}_e$ varies from element to element.

- Global jump Assuming elements e and f share and edge, build the $\mathbb{C}$ matrix as follows:

$$\mathbb{C}_{ef} = -\beta \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma \qquad \text{for } e \neq f \qquad (10.8)$$

  supplemented by

$$\mathbb{C}_{ee} = \beta \sum_{f \neq e} \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma$$

- Local jump Let $\mathcal{M}_e$ be the macroelement that element e is in. Assuming elements e and f share and edge, build the $\mathbb{C}$ matrix as follows:

$$\mathbb{C}_{ef} = -\beta \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma \qquad \text{for } e \neq f \quad \text{and} \quad f \in \mathcal{M}_e \qquad (10.9)$$

  supplemented by

$$\mathbb{C}_{ee} = \beta \sum_{f \neq e, f \in \mathcal{M}_e} \tilde{\mathbb{S}}_{ef} \frac{1}{\Gamma_{ef}} \int_{\Gamma_{ef}} [\![\psi_e]\!] \, [\![\psi_f]\!] d\Gamma$$

- Macroelement stab Let $\mathcal{M}_e$ be the macroelement that element e is in.
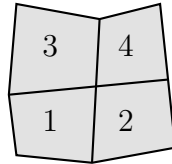
This is less straighforward than the local jump since (for example) elements 1 and 4 do not have an edge in common so that the jump operator cannot be used. One could think of assigning all four elements a single effective viscosity but elements shapes/sizes can differ and the matrix is then not necessarily consistent. One should probably go back to the derivations in Elman, Silvester, and Wathen [394] and see whether a more generic form of the macroelement stabilisation matrix $\mathbb{C}$ could be derived?

10.4.6 The 2D enriched $Q_1^+ \times P_0$ of Fortin

We here consider the enriched $Q_1 \times P_0$ element introduced first by Fortin (1981) [427]. The layout of the degrees of freedom is as follows:
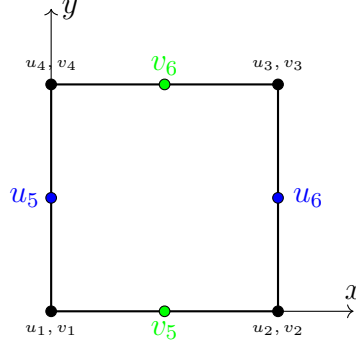
The approximation of the velocity components u and v inside the element is

$$u^h(r,s) = a^u \, \mathcal{N}_1(r,s) + b^u \, \mathcal{N}_2(r,s) + c^u \, \mathcal{N}_3(r,s) + d^u \, \mathcal{N}_4(r,s) + d \, b_5^u(r,s) + e \, b_6^u(r,s)$$

$$v^h(r,s) = a^v \, \mathcal{N}_1(r,s) + b^v \, \mathcal{N}_2(r,s) + c^v \, \mathcal{N}_3(r,s) + d^v \, \mathcal{N}_4(r,s) + d^v b_5^v(r,s) + e^v b_6^v(r,s)$$

where $\mathcal{N}_{1,2,3,4}$ are the standard Q_1 basis functions in 2D and with

$$b_5^u(r,s) = \frac{1}{2}(1-r)(1-s^2) \qquad b_6^u(r,s) = \frac{1}{2}(1+r)(1-s^2)$$

and

$$b_5^v(r,s) = \frac{1}{2}(1-r^2)(1-s) \qquad b_6^v(r,s) = \frac{1}{2}(1-r^2)(1+s)$$

In the end one arrives at

$$
\begin{aligned}
\mathcal{N}_1^u(r,s) &= \mathcal{N}_1(r,s) - \frac{1}{2}b_5^u(r,s) \\
\mathcal{N}_2^u(r,s) &= \mathcal{N}_2(r,s) - \frac{1}{2}b_6^u(r,s) \\
\mathcal{N}_3^u(r,s) &= \mathcal{N}_3(r,s) - \frac{1}{2}b_6^u(r,s) \\
\mathcal{N}_4^u(r,s) &= \mathcal{N}_4(r,s) - \frac{1}{2}b_5^u(r,s) \\
\mathcal{N}_5^u(r,s) &= b_5^u(r,s) \\
\mathcal{N}_6^u(r,s) &= b_6^u(r,s) \\
\\
\mathcal{N}_1^v(r,s) &= \mathcal{N}_1(r,s) - \frac{1}{2}b_5^v(r,s) \\
\mathcal{N}_2^v(r,s) &= \mathcal{N}_2(r,s) - \frac{1}{2}b_5^v(r,s) \\
\mathcal{N}_3^v(r,s) &= \mathcal{N}_3(r,s) - \frac{1}{2}b_6^v(r,s) \\
\mathcal{N}_4^v(r,s) &= \mathcal{N}_4(r,s) - \frac{1}{2}b_6^v(r,s) \\
\mathcal{N}_5^v(r,s) &= b_5^v(r,s) \\
\mathcal{N}_6^v(r,s) &= b_6^v(r,s)
\end{aligned}
\tag{10.10}
$$

We can check for the zero-th order consistency: Let $u(r,s) = C$, then

$$u^h(r,s) = \sum_{i=1}^{6} \mathcal{N}_i^u(r,s) u_i = C \sum_{i=1}^{6} \mathcal{N}_i^u(r,s) = C \sum_{i=1}^{4} \mathcal{N}_i(r,s) = C \tag{10.11}$$

384

10.4.7 The 3D enriched $Q_1^+ \times P_0$ of Fortin

This element is mentioned on p249 of Cuvelier, Segal & van Steenhoven [317]: "The enriched trilinear velocity-constant pressure element is probably the simplest admissible 3D element." Fortin [427] designed a simple LBB-stable Q_1 element to which mid-face nodes are added, i.e. a 'bubble' Q_2 function is added on each face. However, only $\vec{v} \cdot \vec{n}$ is present on these mid-face nodes:
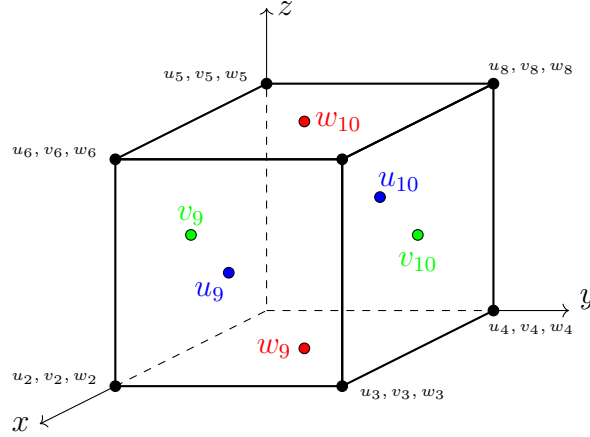
Fortin states: "this element satisfies the B.B. condition and is probably the simplest 3-D element to do So. This unfortunately does not mean that it is more accurate (at least on regular meshes)." and "the element satisfies the B.B. condition. It can therefore be used in a non-regular mesh without fear. The number of degrees of freedom is approximately double with respect to the $Q_1 \times P_0$ element and this is reflected by an increased number of vortices and a reduction of their size. However, there seems to be a qualitative deficiency of these vortices since they do not easily assemble into complex flows. Only numerical experiments can give the final answer." This element is mentioned/used in [1151, 89, 1393].

Considering a single element, we have

- Q_1: $2 \times 2 \times 2 \times 3 = 24$ velocity dofs

- Q_1^+: $2 \times 2 \times 2 \times 3 + 6 = 30$ velocity dofs:

$$\vec{V}^T = (\underbrace{u_1, v_1, w_1, \dots, u_8, v_8, w_8}_{Q_1 \ dofs}, \underbrace{u_9, v_9, w_9, u_{10}, v_{10}, w_{10}}_{bubble \ dofs})$$

  The big difference with all other elements so far is the fact that the dofs u_9, v_9, w_9 are not colocated (same for the other three). u_9 lives in the middle of the $r = -1$ face, v_9 lives in the middle of the $s = -1$ face and w_9 lives in the middle of the $t = -1$ face.

- Q_2: $3 \times 3 \times 3 \times 3 = 81$ velocity dofs

Considering a 3D mesh composed of $nel = nelx \times nely \times nelz$ elements:

- Q_1: the total number of Velocity dofs is $NfemV = (nelx + 1) \times (nely + 1) \times (nelz + 1) \times 3$

- Q_1^+: the total number of nodes is

  $NfemV = (nelx+1) \times (nely+1) \times (nelz+1) \times 3 + (nelx+1) \times nely \times nelz + nelx \times (nely+1) \times nelz + nelx \times ne$

- Q_2: the total number of Velocity dofs is $NfemV = (2nelx + 1) \times (2nely + 1) \times (2nelz + 1) \times 3$

When $nelx = nely = nelz = n >> 1$ then the numbers above converge to $3n^3$, $6n^3$ and $24n^3$ respectively. This means that for large meshes the enriched Q_1 uses twice as many dofs as the standard Q_1 while the Q_2 element uses 8 times more.

x-component of velocity The polynomial representation of the velocity in the element is given by

$$u^h(r, s, t) = a + br + cs + dt + ers + frt + gst + hrst + kb_9(r, s, t) + lb_{10}(r, s, t)$$

where the two bubble functions are:

$$b_9^u(r, s, t) = \frac{1}{2}(1 - r)(1 - s^2)(1 - t^2) \qquad b_{10}^u(r, s, t) = \frac{1}{2}(1 + r)(1 - s^2)(1 - t^2)$$

The coordinates of the u_9 dof is (-1,0,0) and the coordinate of the u_{10} dof is $(1, 0, 0)$. We see that the bubble functions are 1 at their nodes and zero at all other nodes. We can actually use a different basis for $1, r, s, t, rs, rt, st, rst$ and we instead choose the standard Q_1 functions so that u^h becomes:

$$u^h(r, s, t) = aN_1 + bN_2 + cN_3 + dN_4 + eN_5 + fN_6 + gN_7 + hN_8 + kb_9(r, s, t) + lb_{10}(r, s, t)$$

We then must find the set of coefficients $\{a \ldots l\}$ and we will do so by requiring that $u^h(r_i, s_i, t_i) = u_i$ for $i = 1, 10$.

The coordinates of all 10 nodes and the values of basis functions at these locations are:

node #	r	s	t	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	b_9^u	b_{10}^u
1	-1	-1	-1	1	0	0	0	0	0	0	0	0	0
2	+1	-1	-1	0	1	0	0	0	0	0	0	0	0
3	+1	+1	-1	0	0	1	0	0	0	0	0	0	0
4	-1	+1	-1	0	0	0	1	0	0	0	0	0	0
5	-1	-1	+1	0	0	0	0	1	0	0	0	0	0
6	+1	-1	+1	0	0	0	0	0	1	0	0	0	0
7	+1	+1	+1	0	0	0	0	0	0	1	0	0	0
8	-1	+1	+1	0	0	0	0	0	0	0	1	0	0
9	-1	0	0	1/4	0	0	1/4	1/4	0	0	1/4	1	0
10	+1	0	0	0	1/4	1/4	0	0	1/4	1/4	0	0	1

We then have the following ten equations:

$$
\begin{aligned}
u_1 = u^h(r_1, s_1, t_1) &= a \\
u_2 = u^h(r_1, s_1, t_1) &= b \\
u_3 = u^h(r_1, s_1, t_1) &= c \\
u_4 = u^h(r_1, s_1, t_1) &= d \\
u_5 = u^h(r_1, s_1, t_1) &= e \\
u_6 = u^h(r_1, s_1, t_1) &= f \\
u_7 = u^h(r_1, s_1, t_1) &= g \\
u_8 = u^h(r_1, s_1, t_1) &= h \\
u_9 = u^h(r_9, s_9, t_9) &= \frac{1}{4}(a + d + e + h) + k \\
u_{10} = u^h(r_{10}, s_{10}, t_{10}) &= \frac{1}{4}(b + c + f + g) + l
\end{aligned}
$$

or,

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1/4 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1/4 & 1 & 0 \\
0 & 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1
\end{pmatrix}
\cdot
\begin{pmatrix}
a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ k \\ l
\end{pmatrix}
=
\begin{pmatrix}
u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10}
\end{pmatrix}
$$

This yields

$$
\begin{aligned}
a &= u_1 \\
b &= u_2 \\
c &= u_3 \\
d &= u_4 \\
e &= u_5 \\
f &= u_6 \\
g &= u_7 \\
h &= u_8 \\
k &= u_9 - \frac{1}{4}(u_1 + u_4 + u_5 + u_8) \\
l &= u_{10} - \frac{1}{4}(u_2 + u_3 + u_6 + u_7)
\end{aligned}
$$

and then

$$
\begin{aligned}
u^h(r,s,t) &= aN_1 + bN_2 + cN_3 + dN_4 + eN_5 + fN_6 + gN_7 + hN_8 + kb_9^u(r,s,t) + lb_{10}^u(r,s,t) \\
&= u_1N_1 + u_2N_2 + u_3N_3 + u_4N_4 + u_5N_5 + u_6N_6 + u_7N_7 + u_8N_8 \\
&\quad + \left[u_9 - \frac{1}{4}(u_1 + u_4 + u_5 + u_8)\right]b_9^u(r,s,t) + \left[u_{10} - \frac{1}{4}(u_2 + u_3 + u_6 + u_7)\right]b_{10}^u(r,s,t) \\
&= \left(u_1 - \frac{1}{4}b_9\right)N_1 + \left(u_2 - \frac{1}{4}b_{10}\right)N_2 + \left(u_3 - \frac{1}{4}b_{10}\right)N_3 + \left(u_4 - \frac{1}{4}b_9\right)N_4 + \\
&\quad \left(u_5 - \frac{1}{4}b_9\right)N_5 + \left(u_6 - \frac{1}{4}b_{10}\right)N_6 + \left(u_7 - \frac{1}{4}b_{10}\right)N_7 + \left(u_8 - \frac{1}{4}b_9\right)N_8 + \\
&\quad b_9^u(r,s,t)u_9 + b_{10}^u(r,s,t)u_{10}
\end{aligned}
$$

Finally, we can write the basis functions for the u field:

$$
\begin{aligned}
N_1^u(r,s,t) &= N_1(r,s,t) - \frac{1}{4}b_9^u(r,s,t) \\[4pt]
N_2^u(r,s,t) &= N_2(r,s,t) - \frac{1}{4}b_{10}^u(r,s,t) \\[4pt]
N_3^u(r,s,t) &= N_3(r,s,t) - \frac{1}{4}b_{10}^u(r,s,t) \\[4pt]
N_4^u(r,s,t) &= N_4(r,s,t) - \frac{1}{4}b_9^u(r,s,t) \\[4pt]
N_5^u(r,s,t) &= N_5(r,s,t) - \frac{1}{4}b_9^u(r,s,t) \\[4pt]
N_6^u(r,s,t) &= N_6(r,s,t) - \frac{1}{4}b_{10}^u(r,s,t) \\[4pt]
N_7^u(r,s,t) &= N_7(r,s,t) - \frac{1}{4}b_{10}^u(r,s,t) \\[4pt]
N_8^u(r,s,t) &= N_8(r,s,t) - \frac{1}{4}b_9^u(r,s,t) \\[4pt]
N_9^u(r,s,t) &= b_9^u(r,s,t) \\[2pt]
N_{10}^u(r,s,t) &= b_{10}^u(r,s,t)
\end{aligned}
$$

And it is easy to verify that

$$
\sum_{i=1}^{10} N_i^u(r,s,t) = 1 \qquad \forall r,s,t
$$

During the implementation phase we will need the derivatives of the basis functions, which are trivial for the standard Q_1 basis functions N_i. Remain then

$$
\begin{aligned}
\partial_r b_9^u(r,s,t) &= \frac{\partial}{\partial r}\left(\frac{1}{2}(1-r)(1-s^2)(1-t^2)\right) = -\frac{1}{2}(1-s^2)(1-t^2) \\[6pt]
\partial_s b_9^u(r,s,t) &= \frac{\partial}{\partial s}\left(\frac{1}{2}(1-r)(1-s^2)(1-t^2)\right) = -(1-r)s(1-t^2) \\[6pt]
\partial_t b_9^u(r,s,t) &= \frac{\partial}{\partial t}\left(\frac{1}{2}(1-r)(1-s^2)(1-t^2)\right) = -(1-r)(1-s^2)t \\[10pt]
\partial_r b_{10}^u(r,s,t) &= \frac{\partial}{\partial r}\left(\frac{1}{2}(1+r)(1-s^2)(1-t^2)\right) = \frac{1}{2}(1-s^2)(1-t^2) \\[6pt]
\partial_s b_{10}^u(r,s,t) &= \frac{\partial}{\partial s}\left(\frac{1}{2}(1+r)(1-s^2)(1-t^2)\right) = -(1+r)s(1-t^2) \\[6pt]
\partial_t b_{10}^u(r,s,t) &= \frac{\partial}{\partial t}\left(\frac{1}{2}(1+r)(1-s^2)(1-t^2)\right) = -(1+r)(1-s^2)t
\end{aligned}
$$

y-component of velocity The bubbles are given by

$$
b_9^v(r,s,t) = \frac{1}{2}(1-r^2)(1-s)(1-t^2) \qquad b_{10}^v(r,s,t) = \frac{1}{2}(1-r^2)(1+s)(1-t^2)
$$

The coordinates of all 10 nodes and the values of basis functions at these locations are:

node #	r	s	t	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	b_9^v	b_{10}^v
1	-1	-1	-1	1	0	0	0	0	0	0	0	0	0
2	+1	-1	-1	0	1	0	0	0	0	0	0	0	0
3	+1	+1	-1	0	0	1	0	0	0	0	0	0	0
4	-1	+1	-1	0	0	0	1	0	0	0	0	0	0
5	-1	-1	+1	0	0	0	0	1	0	0	0	0	0
6	+1	-1	+1	0	0	0	0	0	1	0	0	0	0
7	+1	+1	+1	0	0	0	0	0	0	1	0	0	0
8	-1	+1	+1	0	0	0	0	0	0	0	1	0	0
9	0	-1	0	1/4	1/4	0	0	1/4	1/4	0	0	1	0
10	0	+1	0	0	0	1/4	1/4	0	0	1/4	1/4	0	1

Then

$$
\begin{aligned}
N_1^v(r,s,t) &= N_1(r,s,t) - \frac{1}{4}b_9^v(r,s,t) \\[6pt]
N_2^v(r,s,t) &= N_2(r,s,t) - \frac{1}{4}b_9^v(r,s,t) \\[6pt]
N_3^v(r,s,t) &= N_3(r,s,t) - \frac{1}{4}b_{10}^v(r,s,t) \\[6pt]
N_4^v(r,s,t) &= N_4(r,s,t) - \frac{1}{4}b_{10}^v(r,s,t) \\[6pt]
N_5^v(r,s,t) &= N_5(r,s,t) - \frac{1}{4}b_9^v(r,s,t) \\[6pt]
N_6^v(r,s,t) &= N_6(r,s,t) - \frac{1}{4}b_9^v(r,s,t) \\[6pt]
N_7^v(r,s,t) &= N_7(r,s,t) - \frac{1}{4}b_{10}^v(r,s,t) \\[6pt]
N_8^v(r,s,t) &= N_8(r,s,t) - \frac{1}{4}b_{10}^v(r,s,t) \\[6pt]
N_9^v(r,s,t) &= b_9^v(r,s,t) \\
N_{10}^v(r,s,t) &= b_{10}^v(r,s,t)
\end{aligned}
\tag{10.12}
$$

z-component of velocity The bubbles are given by

$$
b_9^w(r,s,t) = \frac{1}{2}(1-r^2)(1-s^2)(1-t) \qquad b_{10}^w(r,s,t) = \frac{1}{2}(1-r^2)(1-s^2)(1+t)
$$

The coordinates of all 10 nodes and the values of basis functions at these locations are:

node #	r	s	t	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	b_9^w	b_{10}^w
1	-1	-1	-1	1	0	0	0	0	0	0	0	0	0
2	+1	-1	-1	0	1	0	0	0	0	0	0	0	0
3	+1	+1	-1	0	0	1	0	0	0	0	0	0	0
4	-1	+1	-1	0	0	0	1	0	0	0	0	0	0
5	-1	-1	+1	0	0	0	0	1	0	0	0	0	0
6	+1	-1	+1	0	0	0	0	0	1	0	0	0	0
7	+1	+1	+1	0	0	0	0	0	0	1	0	0	0
8	-1	+1	+1	0	0	0	0	0	0	0	1	0	0
9	0	0	-1	1/4	1/4	1/4	1/4	0	0	0	0	1	0
10	0	0	+1	0	0	0	0	1/4	1/4	1/4	1/4	0	1

$$N_1^w(r,s,t) = N_1(r,s,t) - \frac{1}{4}b_9^w(r,s,t)$$

$$N_2^w(r,s,t) = N_2(r,s,t) - \frac{1}{4}b_9^w(r,s,t)$$

$$N_3^w(r,s,t) = N_3(r,s,t) - \frac{1}{4}b_9^w(r,s,t)$$

$$N_4^w(r,s,t) = N_4(r,s,t) - \frac{1}{4}b_9^w(r,s,t)$$

$$N_5^w(r,s,t) = N_5(r,s,t) - \frac{1}{4}b_{10}^w(r,s,t)$$

$$N_6^w(r,s,t) = N_6(r,s,t) - \frac{1}{4}b_{10}^w(r,s,t)$$

$$N_7^w(r,s,t) = N_7(r,s,t) - \frac{1}{4}b_{10}^w(r,s,t)$$

$$N_8^w(r,s,t) = N_8(r,s,t) - \frac{1}{4}b_{10}^w(r,s,t)$$

$$N_9^w(r,s,t) = b_9^w(r,s,t)$$

$$N_{10}^w(r,s,t) = b_{10}^w(r,s,t)$$

A word about the B matrix   We have

$$u^h(r,s,t) = \sum_{i=1}^{10} N_i^u(r,s,t)u_i \tag{10.13}$$

$$v^h(r,s,t) = \sum_{i=1}^{10} N_i^v(r,s,t)v_i \tag{10.14}$$

$$w^h(r,s,t) = \sum_{i=1}^{10} N_i^w(r,s,t)w_i \tag{10.15}$$

Normally we do not make a distinction between the basis functions associated to u,v,w but because of the bubbles on the faces we now have to.

We have previously established that the strain rate vector $\vec{\dot{\varepsilon}}$ is:

$$\vec{\dot{\varepsilon}} = \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial w}{\partial z} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \\ \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \end{pmatrix} = \begin{pmatrix} \sum_i \frac{\partial N_i^u}{\partial x} u_i \\ \sum_i \frac{\partial N_i^v}{\partial y} v_i \\ \sum_i \frac{\partial N_i^w}{\partial z} w_i \\ \sum_i (\frac{\partial N_i^u}{\partial y} u_i + \frac{\partial N_i^v}{\partial x} v_i) \\ \sum_i (\frac{\partial N_i^u}{\partial z} u_i + \frac{\partial N_i^w}{\partial x} w_i) \\ \sum_i (\frac{\partial N_i^v}{\partial z} v_i + \frac{\partial N_i^w}{\partial y} w_i) \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial N_1^u}{\partial x} & 0 & 0 & \cdots & \frac{\partial N_{10}^u}{\partial x} & 0 & 0 \\ 0 & \frac{\partial N_1^v}{\partial y} & 0 & \cdots & 0 & \frac{\partial N_{10}^v}{\partial y} & 0 \\ 0 & 0 & \frac{\partial N_1^w}{\partial z} & \cdots & 0 & 0 & \frac{\partial N_{10}^w}{\partial z} \\ \frac{\partial N_1^u}{\partial y} & \frac{\partial N_1^v}{\partial x} & 0 & \cdots & \frac{\partial N_{10}^u}{\partial x} & \frac{\partial N_{10}^v}{\partial x} & 0 \\ \frac{\partial N_1^u}{\partial z} & 0 & \frac{\partial N_1^w}{\partial x} & \cdots & \frac{\partial N_{10}^u}{\partial z} & 0 & \frac{\partial N_{m_v}^w}{\partial x} \\ 0 & \frac{\partial N_1^v}{\partial z} & \frac{\partial N_1^w}{\partial y} & \cdots & 0 & \frac{\partial N_{10}^v}{\partial z} & \frac{\partial N_{10}^w}{\partial y} \end{pmatrix}}_{B} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ w_1 \\ u_2 \\ v_2 \\ w_2 \\ u_3 \\ v_3 \\ \cdots \\ u_{10} \\ v_{10} \\ w_{10} \end{pmatrix}}_{\vec{V}}$$

10.4.8 The $Q_1 \times Q_1$-stab
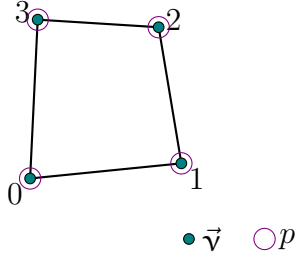
4 vel. nodes, 4 press. nodes   8 vel. nodes, 8 press. nodes
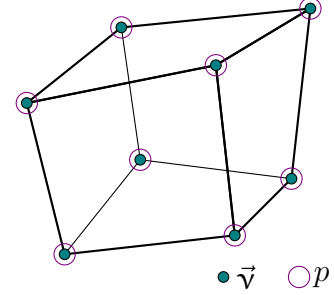
The $\boldsymbol{Q}_1 \times Q_1$ element is not LBB-stable but it can be stabilised. Despite some applications in geodynamics (it is used in Burstedde, Ghattas, Stadler, Tu, and Wilcox [202] (2009) and Burstedde et al. [203] (2013)), it is not appropriate for buoyancy-driven flows, as shown in Thieulot and Bangerth [1345].

See Norburn and Silvester [1001] (2001) for a fourier analysis of the normal and stablised (a la Hughes, Franca, and Balestra [639] (1986)) $\boldsymbol{Q}_1 \times Q_1$ element. Stabilisation is worked out out in Dohrmann and Bochev [358] (2004), Bochev, Dohrmann, and Gunzburger [108] (2006), and Bochev and Dohrmann [109] (2006).

- $\boldsymbol{Q}_1 \times P_0$-stab. Pro: stabilisation can be switched off; Con: stabilisation for deformed elements? problem near boundaries: incomplete stencil? choice of parameter β.

- $\boldsymbol{Q}_1 \times Q_1$-stab. Pro: easier to implement than $\boldsymbol{Q}_1 \times P_0$-stab, stabilisation local to element, easier when elements are not rectangular, no free parameter; Con: stabilisation cannot be switched off.

📚 Relevant Literature: Schneider, Raithby, and Yovanovich [1206], Tezduyar, Mittal, Ray, and Shih [1335], Tezduyar [1333], Gresho, Chan, Christon, and Hindmarsch [514], Idelsohn, Storti, and Nigro [651], Knobloch and Tobiska [757], Franca, Oliveira, and Sarkis [437], and Li, He, and Chen [827]. See Braack and Lube [135] for a review of local projection stabilisation for incompressible flow problems.

This unstable pair is also used in ice sheet modelling Helanow and Ahlkrona [594] , Zhang, Ju, Gunzburger, Ringler, and Price [1506], Zwinger, Greve, Gagliardini, Shiraiwa, and Lyly [1551]. A $\boldsymbol{P}_1 \times P_1$ version of it is used in Karabelas, Haase, Plank, and Augustin [711] (2020).

10.4.9 The $Q_1^R \times P_0$ à la Rannacher-Turek

This element is the natural quadrilateral analogue of the well-known triangular P_1^{nc}Stokes element of Crouzeix-Raviart [309]. This element is sometimes called $\boldsymbol{Q}_1^{rot} \times Q_0$ or the Rannacher-Turek element [688, Section 3.6.5] (see also Appendix B.4, example B.53 of John [688]). This rectangular nonconforming [308] element is termed the rotated $\boldsymbol{Q}_1$ element because of the fact that $r^2 - s^2$ can be generated from rs (occurring in the bilinear Q_1 element) by a rotation of 45° [242, p93]. The velocity approximation is achieved by rotated dim-linear functions that have continuous degrees of freedom on the faces of the mesh cells as we have seen in Section 8.3.9. This element was introduced in Rannacher & Turek (1992) [1112] has been proven to satisfy the inf-sup condition. It has been

studied comprehensively in Schieweck (1997)[10], [1240] and in Turek [1383, 1381]. Superconvergence properties have also been reported [924, 923]. It has been used in 2D [870] and 3D [752, 473] and forms the basis of the FeatFlow software[11]. It is used in the PhD thesis of Gastaldo [468] and Ouazzi [1024]. It has been successfully coupled to multigrid solvers [244, 1385]. This element has been compared to the stabilised $\boldsymbol{Q}_1 \times P_0$ element [833]. It is mentioned in [562]

It essentially comes in two flavours, the Middle Point (MP) and the Mid Value (MV) one.

Remark. *John [688] explains that: "For the point-value-oriented non-conforming finite element spaces (MP), the value of the Dirichlet boundary condition in the barycenter of the faces at the boundary is taken. Using the mean- value-oriented spaces (MV), one computes the integrals of the boundary condition on these faces and normalizes with the area of the faces to set the boundary values. In the case of homogeneous Dirichlet boundary conditions, the boundary values computed in both ways are zero."*

Remark. *John also makes a very important point: "There are also unmapped (non-parametric) versions of these finite element spaces, which define the polynomials directly on the mesh cell K. It is shown in Rannacher and Turek [1112] (1992) that these versions are inf-sup stable on more general meshes than the mapped (parametric) version of the $\boldsymbol{Q}_1^{rot} \times Q_0$ finite element, e.g., on strongly nonuniform meshes. Considering all four types of $\boldsymbol{Q}_1^{rot} \times Q_0$ finite elements, the optimal order of convergence on perturbed meshes is achieved only by the mean- value-oriented version of the unmapped $\boldsymbol{Q}_1^{rot} \times Q_0$ finite element.*

Mahmood *et al.* [870] mention a very important fact:

> The chosen nonconforming element requires additional stabilization for handling the deformation tensor formulation due to missing Korn's inequality (see Horgan [625] (1995), Knobloch [756] (2000), Brenner [153] (2004) for details about Korn's equality). To this end we employ the standard edge oriented stabilization (Turek, Ouazzi, and Schmachtel [1385] (2002), Turek and Ouazzi [1384] (2007)) in our simulations.

This is a rather unfortunate fact that although LBB stable this element needs an additional term in the weak form (see Turek *et al.* (2002) [1385]) so as to suppress parasitic velocity modes when the div-grad formulation of the Stokes equation is used (as opposed to the Laplace formulation – see [363, Section 6.5.2]).

This element is used in Hansbo, Larson, and Larson [563] (2001) in the context of near incompressible elasticity. It is mentioned that it does not fulfill the discrete Korn's inequality. It is then stabilised in a discontinuous Galerkin framework.

We find on page 47 of John [688]:

> It should be noted that estimates of Korn's type do not hold for all spaces, e.g. they do not hold for certain non-conforming finite element spaces.

🔖 Relevant LiteratureSheen [1233] (2020)
https://defelement.com/elements/rannacher-turek.html

10.4.10   The $Q_2 \times Q_0$ pair

Quadratic velocities, constant pressure. The element satisfies the inf-sup condition, but the constant pressure assumption may require fine discretisation. source?

I have implemented it in STONE **??** using the penalty method.

---

[10]Habilitation thesis in German
[11]http://www.featflow.de/en/index.html

10.4.11 The $Q_2 \times (Q_1 + Q_0)$ pair

It is a rather peculiar element pair (triplet?). The velocity space is the standard Q_2 space but the pressure space is the sum of two spaces, i.e. Q_1 *and* Q_0. Please see Section 10.5.11 on the $\boldsymbol{P}_2 \times (P_1 + P_0)$ element.
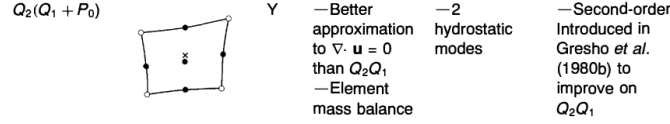
Taken from Gresho and Sani [517]'s book.

It is implemented in STONE 120.

10.4.12 The $Q_2^+ \times Q_{-1}$ pair

This element pair is mentioned in Carey and Oden [224] (1986). One can introduce the bubble function $b(r,s) = (1-r^2)(1-s^2)$ and consider velocity approximations of the form

$$u_h(r,s) = \sum_{i=1}^{9} \mathcal{N}_i(r,s)u_i + a_1 s b(r,s)$$

$$v_h(r,s) = \sum_{i=1}^{9} \mathcal{N}_i(r,s)v_i + a_2 r b(r,s)$$

Using these expressions we can obtain a relationship between a_1 and a_2 such that the element is stable.

10.4.13 The DSSY pair in 2D

This element is often referred to as the 'DSSY' element because of the four authors of the original paper: Douglas, Santos, sheen and Ye (1999) [367]. The non-conforming finite element space Q_l is defined based on the reference square element on $[-1, 1]^2$:

$$Q_l = \text{Span} \{1, r, s, \theta_l(r) - \theta_l(s)\} \qquad l = 1, \text{ or } 2$$

with

$$
\begin{aligned}
\theta_1(r) &= r^2 - \frac{5}{3}r^4 \\
\theta_1'(r) &= 2r - \frac{20}{3}r^3 \\
\theta_2(r) &= r^2 - \frac{25}{6}r^4 + \frac{7}{2}r^6 & (10.16) \\
\theta_2'(r) &= 2r - \frac{50}{3}r^3 + 21r^5 & (10.17)
\end{aligned}
$$

The dimension of Q_l is four and the θ_l functions look like:
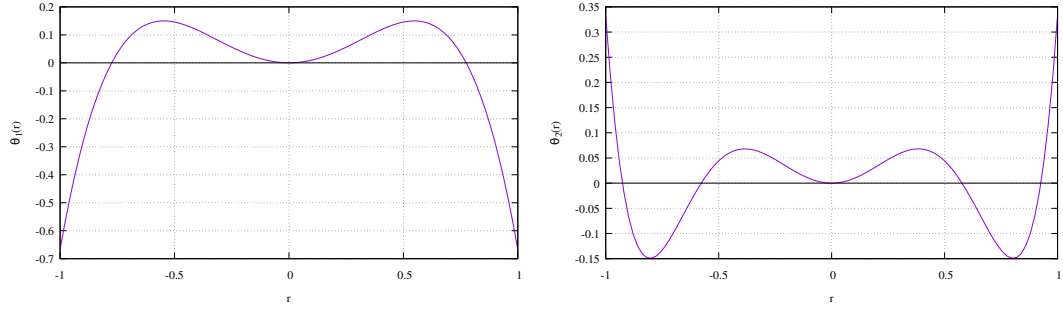
We have:

- $\theta_1(r = -1) = \theta_1(r = +1) = -\frac{2}{3}$, $\theta_1(r = 0) = 0$

- $\theta_2(r = -1) = \theta_2(r = +1) = \frac{1}{3}$, $\theta_2(r = 0) = 0$

The nodes are situated at the mid-edges of the quadrilateral:
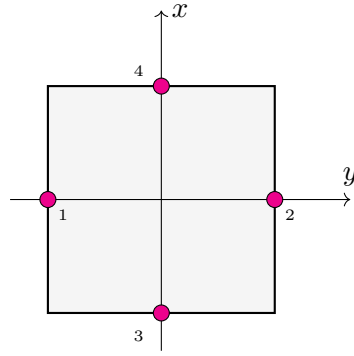
The basis function corresponding to the node $(1, 0)$ is given by

$$
\begin{aligned}
\mathcal{N}_1(r, s)^{(l)} &= \frac{1}{4} - \frac{1}{2}r + \frac{\theta_l(r) - \theta_l(s)}{4\theta_l(1)} \\
\mathcal{N}_2(r, s)^{(l)} &= \frac{1}{4} + \frac{1}{2}r + \frac{\theta_l(r) - \theta_l(s)}{4\theta_l(1)} \\
\mathcal{N}_3(r, s)^{(l)} &= \frac{1}{4} - \frac{1}{2}s - \frac{\theta_l(r) - \theta_l(s)}{4\theta_l(1)} \\
\mathcal{N}_4(r, s)^{(l)} &= \frac{1}{4} + \frac{1}{2}s - \frac{\theta_l(r) - \theta_l(s)}{4\theta_l(1)}
\end{aligned}
\tag{10.18}
$$

We can easily verify that $\sum_i \mathcal{N}_i(r, s, t) = 1$ and that $\mathcal{N}_i(\vec{r}_j) = \delta_{ij}$:

$$\mathcal{N}_1^{(l)}(r_1, s_1) = \frac{1}{4} - \frac{1}{2}(-1) + \frac{\theta_l(-1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} + \frac{1}{2} + \frac{\theta_l(-1)}{4\theta_l(1)} = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

$$\mathcal{N}_1^{(l)}(r_2, s_2) = \frac{1}{4} - \frac{1}{2}(+1) + \frac{\theta_l(+1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{2} + \frac{\theta_l(+1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{2} + \frac{1}{4} = 0$$

$$\mathcal{N}_1^{(l)}(r_3, s_3) = \frac{1}{4} - \frac{1}{2}(0) + \frac{\theta_l(0) - \theta_l(-1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_1^{(l)}(r_4, s_4) = \frac{1}{4} - \frac{1}{2}(0) + \frac{\theta_l(0) - \theta_l(+1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_2^{(l)}(r_1, s_1) = \frac{1}{4} + \frac{1}{2}(-1) + \frac{\theta_l(-1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{2} + \frac{1}{4} = 0$$

$$\mathcal{N}_2^{(l)}(r_2, s_2) = \frac{1}{4} + \frac{1}{2}(+1) + \frac{\theta_l(+1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

$$\mathcal{N}_2^{(l)}(r_3, s_3) = \frac{1}{4} + \frac{1}{2}(0) + \frac{\theta_l(0) - \theta_l(-1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_2^{(l)}(r_4, s_4) = \frac{1}{4} + \frac{1}{2}(0) + \frac{\theta_l(0) - \theta_l(+1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_3^{(l)}(r_1, s_1) = \frac{1}{4} - \frac{1}{2}(0) - \frac{\theta_l(-1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_3^{(l)}(r_2, s_2) = \frac{1}{4} - \frac{1}{2}(0) - \frac{\theta_l(+1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_3^{(l)}(r_3, s_3) = \frac{1}{4} - \frac{1}{2}(-1) - \frac{\theta_l(0) - \theta_l(-1)}{4\theta_l(1)} = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

$$\mathcal{N}_3^{(l)}(r_4, s_4) = \frac{1}{4} - \frac{1}{2}(+1) - \frac{\theta_l(0) - \theta_l(+1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{2} + \frac{1}{4} = 0$$

$$\mathcal{N}_4^{(l)}(r_1, s_1) = \frac{1}{4} + \frac{1}{2}(0) - \frac{\theta_l(-1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_4^{(l)}(r_2, s_2) = \frac{1}{4} + \frac{1}{2}(0) - \frac{\theta_l(+1) - \theta_l(0)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{4} = 0$$

$$\mathcal{N}_4^{(l)}(r_3, s_3) = \frac{1}{4} + \frac{1}{2}(-1) - \frac{\theta_l(0) - \theta_l(-1)}{4\theta_l(1)} = \frac{1}{4} - \frac{1}{2} + \frac{1}{4} = 0$$

$$\mathcal{N}_4^{(l)}(r_4, s_4) = \frac{1}{4} + \frac{1}{2}(1) - \frac{\theta_l(0) - \theta_l(1)}{4\theta_l(1)} = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

The basis functions can also be explicitly written for θ_1 as in Cai *et al.* [219]:

$$\mathcal{N}_1(r, s)^{(l)} = \frac{1}{4} - \frac{1}{2}r - \frac{3}{8}\left[\left(r^2 - \frac{5}{3}r^4\right) - \left(s^2 - \frac{5}{3}s^4\right)\right]$$

$$\mathcal{N}_2(r, s)^{(l)} = \frac{1}{4} + \frac{1}{2}r - \frac{3}{8}\left[\left(r^2 - \frac{5}{3}r^4\right) - \left(s^2 - \frac{5}{3}s^4\right)\right]$$

$$\mathcal{N}_3(r, s)^{(l)} = \frac{1}{4} - \frac{1}{2}s + \frac{3}{8}\left[\left(r^2 - \frac{5}{3}r^4\right) - \left(s^2 - \frac{5}{3}s^4\right)\right]$$

$$\mathcal{N}_4(r, s)^{(l)} = \frac{1}{4} + \frac{1}{2}s + \frac{3}{8}\left[\left(r^2 - \frac{5}{3}r^4\right) - \left(s^2 - \frac{5}{3}s^4\right)\right] \tag{10.19}$$

The derivatives of the basis functions are as follows:

$$\partial_r \mathcal{N}_1(r,s)^{(l)} = -\frac{1}{2} + \frac{\theta_l'(r)}{4\theta_l(1)}$$

$$\partial_r \mathcal{N}_2(r,s)^{(l)} = +\frac{1}{2} + \frac{\theta_l'(r)}{4\theta_l(1)}$$

$$\partial_r \mathcal{N}_3(r,s)^{(l)} = -\frac{\theta_l'(r)}{4\theta_l(1)}$$

$$\partial_r \mathcal{N}_4(r,s)^{(l)} = -\frac{\theta_l'(r)}{4\theta_l(1)} \tag{10.20}$$

$$\partial_s \mathcal{N}_1(r,s)^{(l)} = -\frac{\theta_l'(s)}{4\theta_l(1)}$$

$$\partial_s \mathcal{N}_2(r,s)^{(l)} = -\frac{\theta_l'(s)}{4\theta_l(1)}$$

$$\partial_s \mathcal{N}_3(r,s)^{(l)} = -\frac{1}{2} + \frac{\theta_l'(s)}{4\theta_l(1)}$$

$$\partial_s \mathcal{N}_4(r,s)^{(l)} = +\frac{1}{2} + \frac{\theta_l'(s)}{4\theta_l(1)} \tag{10.21}$$

Note that a correction was issued in Cai, Douglas Jr, Santos, Sheen, and Ye [218] (2000) if a true quadrilateral (i.e., one having two opposite, nonparallel edges) is included in the partition. The authors state that in the case of rectangles the original method is fine.

🔖 Relevant Literature: Park & Sheen (2003) [1034], Jeon *et al.* (2013) [683], Park, Sheen & Shin (2013) [1035], Bangerth *et al.* (2017) [47], Sheen (2020) [1233]

10.4.14 The Han pair

It is based on Han [559] (also mentioned in Sheen (2020) [1233]). The nodes are at the same location as for the RT element above, but there is an additional bubble function in the middle:
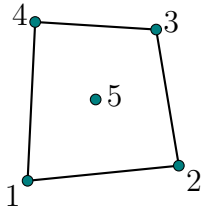
Inside the reference element we assume that a field f can be represented by

$$f^h(r,s) = a + br + cs + d\underbrace{\frac{5r^4 - 3r^2}{2}}_{\phi(r)} + e\underbrace{\frac{5s^4 - 3s^2}{2}}_{\phi(s)}$$

396

We then must have

$$
\begin{aligned}
f_1 &= f^h(r = 1, s = 0) & &= a + b + d \\
f_2 &= f^h(r = 0, s = 1) & &= a + c + e \\
f_3 &= f^h(r = -1, s = 0) & &= a - b + d \\
f_4 &= f^h(r = 0, s = -1) & &= a - c + e \\
f_5 &= f^h(r = 0, s = 0) & &= a
\end{aligned}
$$

and we easily get

$$
a = f_5 \qquad f_1 - f_3 = 2b \qquad f_2 - f_4 = 2c
$$

followed by

$$
d = f_1 - a - b = f_1 - f_5 - \frac{1}{2}(f_1 - f_3) = \frac{f_1 - 2f_5 + f_3}{2}
$$

and

$$
e = f_2 - a - c = f_2 - f_5 - \frac{1}{2}(f_2 - f_4) = \frac{f_2 - 2f_5 + f_4}{2}
$$

Finally:

$$
f(r, s) = f_5 + \frac{1}{2}(f_1 - f_3)r + \frac{1}{2}(f_2 - f_4)s + \frac{f_1 - 2f_5 + f_3}{2}\phi(r) + \frac{f_2 - 2f_5 + f_4}{2}\phi(s)
$$

i.e.

$$
f(r, s) = \left(\frac{r + \phi(r)}{2} \right) f_1 + \left(\frac{s + \phi(s)}{2} \right) f_2 + \left(-\frac{r - \phi(r)}{2} \right) f_3 + \left(-\frac{s - \phi(s)}{2} \right) f_4 + (1 - \phi(r) - \phi(s))\, f_5
$$

which has us define

$$
\begin{aligned}
\mathcal{N}_1(r, s) &= \frac{r + \phi(r)}{2} \\
\mathcal{N}_2(r, s) &= \frac{s + \phi(s)}{2} \\
\mathcal{N}_3(r, s) &= -\frac{r - \phi(r)}{2} \\
\mathcal{N}_4(r, s) &= -\frac{s - \phi(s)}{2} \\
\mathcal{N}_5(r, s) &= 1 - \phi(r) - \phi(s)
\end{aligned}
$$

We have of course the following properties $\sum_{i=1}^{5} \mathcal{N}_i(r, s) = 1$ and $\mathcal{N}_i(r_j, s_j) = \delta_{ij}, \; i, j \in 1, 5$. The

partial derivatives of the basis functions are as follows

$$
\begin{aligned}
\partial_r \mathcal{N}_1(r,s) &= \frac{1+\phi'(r)}{2} \\
\partial_r \mathcal{N}_2(r,s) &= 0 \\
\partial_r \mathcal{N}_3(r,s) &= -\frac{1-\phi'(r)}{2} \\
\partial_r \mathcal{N}_4(r,s) &= 0 \\
\partial_r \mathcal{N}_5(r,s) &= -\phi'(r) \\
\partial_s \mathcal{N}_1(r,s) &= 0 \\
\partial_s \mathcal{N}_2(r,s) &= \frac{1+\phi'(s)}{2} \\
\partial_s \mathcal{N}_3(r,s) &= 0 \\
\partial_s \mathcal{N}_4(r,s) &= -\frac{1-\phi'(s)}{2} \\
\partial_s \mathcal{N}_5(r,s) &= -\phi'(s)
\end{aligned}
$$

This element is implemented in the `stone_han.py` file in STONE 77 and also in STONE 120.

10.4.15 The Chen nonconforming $Q_1 \times Q_0$ pair (?)

What follows is tentative!
This space is proposed in Chen [243] (1993), albeit not in the context of the Stokes equations.
It is based on the mid-point variant of the RT basis functions,

$$
\begin{aligned}
\mathcal{N}_1(r,s) &= \frac{1}{4}(1 - 2s - (r^2 - s^2)) \\
\mathcal{N}_2(r,s) &= \frac{1}{4}(1 + 2r + (r^2 - s^2)) \\
\mathcal{N}_3(r,s) &= \frac{1}{4}(1 + 2s - (r^2 - s^2)) \\
\mathcal{N}_4(r,s) &= \frac{1}{4}(1 - 2r + (r^2 - s^2))
\end{aligned}
$$

to which a P_2 bubble is added

$$
\phi(r,s) = 1 - \frac{3}{4}(r^2 + s^2)
$$

Note thath this function is zero at locations $\pm 1/\sqrt{3}$ on all four edges and exactly 1 in the middle.
A field f is represented inside the element by

$$
f^h(r,s) = a\mathcal{N}_1(r,s) + b\mathcal{N}_2(r,s) + c\mathcal{N}_3(r,s) + d\mathcal{N}_4(r,s) + e\phi(r,s)
$$

We immediately see that this space is not interpolatory, i.e. the basis function $\phi(r,s)$ cannot be 1 in the middle and 0 at the other four nodes.
Chen [242] also extends this to 3D in the paper.
This space is used for velocity and a Q_0 space is used for pressure in STONE 120 (only because the basis functions above are based on the Rannacher-Turek ones).

10.4.16 The $Q_{k+1,k} \times Q_{k,k+1} \times Q_k^-$

This element pair is introduced in Zhang [1508] (2009) where it is shown to be stable for $k \geq 2$ (along with optimal order of convergence). Then in Huang and Zhang [633] (2011) it is shown to be stable for $k \geq 1$. Note that in both cases the pressure space is the space of discontinuous polynomials fo separated degree k or less, *with spurious modes filtered*. To be precise, Q_k^- is the divergence of the discrete velocity space $Q_{k+1,k} \times Q_{k,k+1}$. The author states:

> This is the first divergence-free element found on nontriangular grids.

We also read

> The maximal possible (in space dimension) discrete pressure space is the divergence of the discrete velocity space. If so, the discrete solution for the velocity would be also divergence-free. [...] to find a divergence-free element on rectangular grids, we seek a finite element space whose divergence would be on to the discontinuous Q_k space. This leads to the new element, approximating the velocity by continuous $Q_{k+1,k} \times Q_{k,k+1}$ polynomials while approximating the pressure by discontinuous polynomials $Q_k^- = div(Q_{k+1,k} \times Q_{k,k+1})$ (see eq(2.4) below for the precise definition where $Q_k^- = P_h$). We note that, by choosing $div(Q_{k+1,k} \times Q_{k,k+1})$ as the discrete finite element space for the pressure, the spurious modes in discontinuous Q_k space are filtered out automatically. On the other side, the discrete velocity is divergence-free if and only if the discrete pressure space is the divergence of the discrete velocity space. Because the divergence of a $Q_{k+1,k} \times Q_{k,k+1}$ polynomial is a Q_k polynomial, the new element makes a perfect match for the two discrete spaces and produces divergence-free solutions for the velocity.

In the 2011 paper we find the following enlightening figure for $Q_{21} \times Q_{12} \times Q_1^-$:
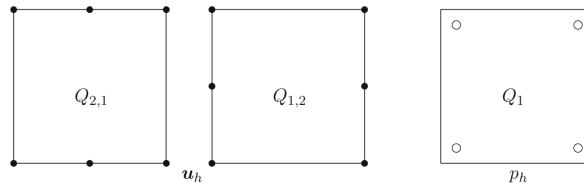
and we read

> The velocity space is the continuous $Q_{2,1} \times Q_{1,2}$ space, i.e., the first component of velocity is a piecewise polynomial of degree 2 in x direction but degree 1 in y direction while the second component is a rotation of the first one. The pressure space is the subspace of discontinuous bilinear polynomials, the divergence of the velocity, to be explicit. [...] we decrease the space Q_1^{dc} for the pressure to Q_1' , by removing all spurious modes, i.e., eliminating one degree of freedom at each vertex.

I do not understand the second statement (yet).

The u-space basis functions are given by

$$\vec{\mathcal{N}}_u = Q_2 \times Q_1 = \begin{pmatrix} \frac{1}{2}r(r-1) \\ 1-r^2 \\ \frac{1}{2}r(r+1) \end{pmatrix} \times \begin{pmatrix} \frac{1}{2}(1-s) & \frac{1}{2}(1+s) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}r(r-1)\frac{1}{2}(1-s) \\ (1-r^2)\frac{1}{2}(1-s) \\ \frac{1}{2}r(r+1)\frac{1}{2}(1-s) \\ \frac{1}{2}r(r-1)\frac{1}{2}(1+s) \\ (1-r^2)\frac{1}{2}(1+s) \\ \frac{1}{2}r(r+1)\frac{1}{2}(1+s) \end{pmatrix} = \begin{pmatrix} \mathcal{N}_1^u \\ \mathcal{N}_2^u \\ \mathcal{N}_3^u \\ \mathcal{N}_4^u \\ \mathcal{N}_5^u \\ \mathcal{N}_6^u \end{pmatrix}$$

$$\vec{\mathcal{N}}_v = Q_1 \times Q_2 = \begin{pmatrix} \frac{1}{2}(1-r) \\ \frac{1}{2}(1+r) \end{pmatrix} \times \begin{pmatrix} \frac{1}{2}s(s-1) & 1-s^2 & \frac{1}{2}s(s+1) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(1-r)\frac{1}{2}s(s-1) \\ \frac{1}{2}(1+r)\frac{1}{2}s(s-1) \\ \frac{1}{2}(1-r)(1-s^2) \\ \frac{1}{2}(1+r)(1-s^2) \\ \frac{1}{2}(1-r)\frac{1}{2}s(s+1) \\ \frac{1}{2}(1+r)\frac{1}{2}s(s+1) \end{pmatrix} = \begin{pmatrix} \mathcal{N}_1^v \\ \mathcal{N}_2^v \\ \mathcal{N}_3^v \\ \mathcal{N}_4^v \\ \mathcal{N}_5^v \\ \mathcal{N}_6^v \end{pmatrix}$$

The u-dofs and v-dofs are living on different nodes. Based on the vectors above the numbering is then

```
4===5===6    5=======6    4=======3
|       |    |       |    |       |
|       |    3       4    |       |
|       |    |       |    |       |
1===2===3    1=======2    1=======2

 u dofs       v dofs       p dofs
```

In total, there are 6 u-dofs, 6 v-dofs and 4 p-dofs per element.

The solution strategy is also quite peculiar (and identical to both papers, but following quote taken from [633]):

> We have to emphasize that the discrete pressure space is introduced for the analysis, but not in the computation. By an iterated penalty method, we obtain the discrete solutions for the pressure without coding the pressure element, as the pressure solution is a byproduct, cf. [30] and Section 4 below. This is an advantage of the divergence-free element, which is not carried to other mixed-finite elements, neither to nonconforming (discontinuous) locally-divergence-free elements [6,10,12,14,15,17,28].

Reading further, we find:

> It is apparent that the discrete velocity solution is divergence-free if and only if the discrete pressure finite element space is the divergence of the discrete velocity finite element space. We note that by (2.4), P_h is a subspace of discontinuous, piecewise bilinear polynomials. As singular vertices are present (see [23,24]), P_h is a proper subset of the discontinuous piecewise Q_1 polynomials. It is possible, but difficult to find a local basis for P_h. However, on the other side, it is the special interest of the divergence-free finite element method that the space P_h can be omitted in computation and the discrete solutions approximating the pressure function in the Stokes equations can be obtained as byproducts, if an iterated penalty method is adopted to solve system (2.5).
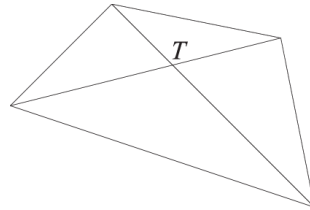
"A 2D vertex of a triangulation is singular if all edges meeting at the vertex form two cross lines [1513]

The iterated penalty method is described (for now) in STONE 161 where the second order version of this element is implemented.

Turning now to $Q_{3,2} \times Q_{2,3} \times Q_2$

```
9=10=11=12    10=11==12   7===8===9
|       |     |      |    |       |
|       |     7   8  9    |       |
5==6==7==8    |      |    4   5   6
|       |     4   5  6    |       |
|       |     |      |    |       |
1==2==3==4    1===2===3   1===2===3


  u dofs         v dofs        p dofs
```

In total, there are 12 u-dofs, 12 v-dofs and 9 p-dofs per element.

The u-space basis functions are given by

$$\vec{\mathcal{N}}_u = Q_3 \times Q_2 = \begin{pmatrix} \frac{1}{16}(-1+r+9r^2-9r^3) \\ \frac{1}{16}(9-27r-9r^2+27r^3) \\ \frac{1}{16}(9+27r-9r^2-27r^3) \\ \frac{1}{16}(-1-r+9r^2+9r^3) \end{pmatrix} \times \begin{pmatrix} \frac{1}{2}s(s-1) & 1-s^2 & \frac{1}{2}s(s+1) \end{pmatrix}$$

The other basis functions follow simply from this. Same for $Q_{4,3}$ and $Q_{3,4}$.

Open questions:

- 3D is discussed in [1508] (2009) for $k \geq 2$ but quid of 3D for $k = 1$? (not discussed in 2011 paper)

- quid of deformed elements?

10.4.17 The $Q_1^{++} \times Q_1$ of Karabelas et al (2020) - 3D

This element is implemented in STONE 82. The two bubble functions are given in Karabelas *et al.* (2020) [711]:

$$b_9(r,s,t) = \left(\frac{27}{32}\right)^3 (1-r^2)(1-s^2)(1-t^2) \cdot (1-r)(1-s)(1-t) = \beta_9(r) \cdot \beta_9(s) \cdot \beta_9(t)$$

$$b_{10}(r,s,t) = \left(\frac{27}{32}\right)^3 (1-r^2)(1-s^2)(1-t^2) \cdot (1+r)(1+s)(1+t) = \beta_{10}(r) \cdot \beta_{10}(s) \cdot \beta_{10}(t)$$

where I have chosen nodes 1 ($\vec{r}_1 = (-1,-1,-1)$) and 7 ($\vec{r}_7 = (+1,+1,+1)$) as diagonally opposed nodes (a requirement from the paper), and with

$$\beta_9(x) = \frac{27}{32}(1-x^2)(1-x) \qquad \beta_{10}(x) = \frac{27}{32}(1-x^2)(1+x)$$

I have added the $(27/32)^3$ coefficients so that these functions are exactly 1 a their corresponding nodes. The term $(1-r^2)(1-s^2)(1-t^2)$ makes sure that the two bubbles are conforming and exactly zero on the 6 faces of the element. In what follows $\tilde{\mathcal{N}}_{1..8}$ are the standard Q_1 basis functions.

401
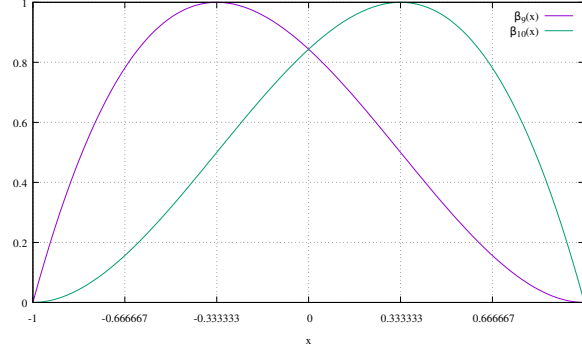
Representation of bubbles $\beta_9(x)$ and $\beta_{10}(x)$

Remark. *Bubble function 9 is not zero at node 10 and vice versa!*

The authors state: "This also allows for a straightforward inclusion in combination with existing finite element codes since all required implementations are purely on the element level". This is especially true if static condensation is used (the authors explain static condensation for the bubbles in the appendix of the paper).

The ten nodes are the standard 8 corners of the Q_1 element as well as $\vec{r}_9 = (-1/3, -1/3, -1/3)$ for b_9 and $\vec{r}_{10} = (1/3, 1/3, 1/3)$ for b_{10}. We have the following approximation of function f inside the element:

$$f^h(r, s, t) = \sum_{i=1}^{8} a_i \tilde{\mathcal{N}}_i(r, s, t) + a_9 b_9(r, s, t) + a_{10} b_{10}(r, s, t)$$

We notice that bubble functions are exactly zero at the corners of the reference element and we can compute the values of the ten polynomials $(\tilde{N}_{1-8}(r, s, t), b_9(r, s, t), b_{10}(r, s, t))$ at the ten nodes:

	$\tilde{\mathcal{N}}_1$	$\tilde{\mathcal{N}}_2$	$\tilde{\mathcal{N}}_3$	$\tilde{\mathcal{N}}_4$	$\tilde{\mathcal{N}}_5$	$\tilde{\mathcal{N}}_6$	$\tilde{\mathcal{N}}_7$	$\tilde{\mathcal{N}}_8$	b_9	b_{10}
$\vec{r}_1 = (-1, -1, -1)$	1	0	0	0	0	0	0	0	0	0
$\vec{r}_2 = (+1, -1, -1)$	0	1	0	0	0	0	0	0	0	0
$\vec{r}_3 = (+1, +1, -1)$	0	0	1	0	0	0	0	0	0	0
$\vec{r}_4 = (-1, +1, -1)$	0	0	0	1	0	0	0	0	0	0
$\vec{r}_5 = (-1, -1, +1)$	0	0	0	0	1	0	0	0	0	0
$\vec{r}_6 = (+1, -1, +1)$	0	0	0	0	0	1	0	0	0	0
$\vec{r}_7 = (+1, +1, +1)$	0	0	0	0	0	0	1	0	0	0
$\vec{r}_8 = (-1, +1, +1)$	0	0	0	0	0	0	0	1	0	0
$\vec{r}_9 = (-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$	8/27	4/27	2/27	4/27	4/27	2/27	1/27	2/27	1	1/8
$\vec{r}_{10} = (+\frac{1}{3}, +\frac{1}{3}, +\frac{1}{3})$	1/27	2/17	4/27	2/27	2/27	4/27	8/27	4/27	1/8	1

We then require that the polynomial representation of f^h of f inside the element is such that

$f^h(\vec{r}_i) = f_i$, i.e.:

$$
\begin{aligned}
f_1 = f^h(r_1, s_1, t_1) &= a_1 \\
f_2 = f^h(r_2, s_1, t_1) &= a_2 \\
f_3 = f^h(r_3, s_1, t_1) &= a_3 \\
f_4 = f^h(r_4, s_1, t_1) &= a_4 \\
f_5 = f^h(r_5, s_1, t_1) &= a_5 \\
f_6 = f^h(r_6, s_1, t_1) &= a_6 \\
f_7 = f^h(r_7, s_1, t_1) &= a_7 \\
f_8 = f^h(r_8, s_1, t_1) &= a_8 \\
f_9 = f^h(r_9, s_9, t_9) &= \frac{1}{27}(8a_1 + 4a_2 + 2a_3 + 4a_4 + 4a_5 + 2a_6 + a_7 + 2a_8) + a_9 + a_{10}/8 \\
f_{10} = f^h(r_{10}, s_{10}, t_{10}) &= \frac{1}{27}(a_1 + 2a_2 + 4a_3 + 2a_4 + 2a_5 + 4a_6 + 8a_7 + 4a_8) + a_9/8 + a_{10}
\end{aligned}
$$

or,

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
8/27 & 4/27 & 2/27 & 4/27 & 4/27 & 2/27 & 1/27 & 2/27 & 1 & 1/8 \\
1/27 & 2/17 & 4/27 & 2/27 & 2/27 & 4/27 & 8/27 & 4/27 & 1/8 & 1
\end{pmatrix}
\cdot
\begin{pmatrix}
a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10}
\end{pmatrix}
=
\begin{pmatrix}
f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \\ f_{10}
\end{pmatrix}
$$

which yields $a_i = f_i$ for $i = 1, ...8$ and

$$
a_9 + a_{10}/8 = \underbrace{f_9 - \frac{1}{27}(8f_1 + 4f_2 + 2f_3 + 4f_4 + 4f_5 + 2f_6 + f_7 + 2f_8)}_{\tilde{f}_9}
$$

$$
a_9/8 + a_{10} = \underbrace{f_{10} - \frac{1}{27}(f_1 + 2f_2 + 4f_3 + 2f_4 + 2f_5 + 4f_6 + 8f_7 + 4f_8)}_{\tilde{f}_{10}}
$$

$$
\begin{aligned}
8a_9 + a_{10} &= 8f_9 + 8\tilde{f}_9 \\
a_9 + 8a_{10} &= 8f_{10} + 8\tilde{f}_{10}
\end{aligned}
$$

and then

$$a_9 = \frac{1}{63}(64f_9 - 8f_{10}) + \frac{1}{63}(64\tilde{f}_9 - 8\tilde{f}_{10})$$

$$= \frac{8}{63}(8f_9 - f_{10}) + \frac{8}{63}(8\tilde{f}_9 - \tilde{f}_{10})$$

$$= \frac{8}{63}(8f_9 - f_{10}) - \frac{1}{27}\frac{8}{63}[8(8f_1 + 4f_2 + 2f_3 + 4f_4 + 4f_5 + 2f_6 + f_7 + 2f_8)$$
$$-(f_1 + 2f_2 + 4f_3 + 2f_4 + 2f_5 + 4f_6 + 8f_7 + 4f_8)]$$

$$= \frac{8}{63}(8f_9 - f_{10}) - \frac{1}{27}\frac{8}{63}(63f_1 + 30f_2 + 12f_3 + 30f_4 + 30f_5 + 12f_6 + 12f_8)$$

$$a_{10} = \frac{1}{63}(64f_{10} - 8f_9) + \frac{1}{63}(64\tilde{f}_{10} - 8\tilde{f}_9)$$

$$= \frac{8}{63}(8f_{10} - f_9) + \frac{8}{63}(8\tilde{f}_{10} - \tilde{f}_9)$$

$$= \frac{8}{63}(8f_{10} - f_9) - \frac{1}{27}\frac{8}{63}[8(f_1 + 2f_2 + 4f_3 + 2f_4 + 2f_5 + 4f_6 + 8f_7 + 4f_8)$$
$$-(8f_1 + 4f_2 + 2f_3 + 4f_4 + 4f_5 + 2f_6 + f_7 + 2f_8)]$$

$$= \frac{8}{63}(8f_{10} - f_9) - \frac{1}{27}\frac{8}{63}(12f_2 + 30f_3 + 12f_4 + 12f_5 + 30f_6 + 63f_7 + 30f_8)$$

We can then write

$$
\begin{aligned}
f^h(r,s,t) &= f_1\mathcal{N}_1(r,s,t) + f_2\mathcal{N}_2(r,s,t) + f_3\mathcal{N}_3(r,s,t) + f_4\mathcal{N}_4(r,s,t)\\
&+ f_5\mathcal{N}_5(r,s,t) + f_6\mathcal{N}_6(r,s,t) + f_7\mathcal{N}_7(r,s,t) + f_8\mathcal{N}_8(r,s,t)\\
&+ \left[\frac{8}{63}(8f_9 - f_{10}) - \frac{1}{27}\frac{8}{63}(63f_1 + 30f_2 + 12f_3 + 30f_4 + 30f_5 + 12f_6 + 12f_8)\right]b_9(r,s,t)\\
&\quad \left[\frac{8}{63}(8f_{10} - f_9) - \frac{1}{27}\frac{8}{63}(12f_2 + 30f_3 + 12f_4 + 12f_5 + 30f_6 + 63f_7 + 30f_8)\right]b_{10}(r,s,t)\\
&= \left(\mathcal{N}_1(r,s,t) - \frac{2^3}{3^3}b_9(r,s,t)\right)f_1\\
&+ \left(\mathcal{N}_2(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)\right)f_2\\
&+ \left(\mathcal{N}_3(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)\right)f_3\\
&+ \left(\mathcal{N}_4(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)\right)f_4\\
&+ \left(\mathcal{N}_5(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)\right)f_5\\
&+ \left(\mathcal{N}_6(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)\right)f_6\\
&+ \left(\mathcal{N}_7(r,s,t) - \frac{2^3}{3^3}b_{10}(r,s,t)\right)f_7\\
&+ \left(\mathcal{N}_8(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)\right)f_8\\
&+ \left(\frac{64}{63}b_9(r,s,t) - \frac{8}{63}b_{10}(r,s,t)\right)f_9 + \left(-\frac{8}{63}b_9(r,s,t) + \frac{64}{63}b_{10}(r,s,t)\right)f_{10} \qquad (10.22)
\end{aligned}
$$

and finally arrive at the basis functions:

$$\mathcal{N}_1(r,s,t) = \tilde{\mathcal{N}}_1(r,s,t) - \frac{2^3}{3^3}b_9(r,s,t)$$

$$\mathcal{N}_2(r,s,t) = \tilde{\mathcal{N}}_2(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)$$

$$\mathcal{N}_3(r,s,t) = \tilde{\mathcal{N}}_3(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)$$

$$\mathcal{N}_4(r,s,t) = \tilde{\mathcal{N}}_4(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}r,s,t)$$

$$\mathcal{N}_5(r,s,t) = \tilde{\mathcal{N}}_5(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)$$

$$\mathcal{N}_6(r,s,t) = \tilde{\mathcal{N}}_6(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)$$

$$\mathcal{N}_7(r,s,t) = \tilde{\mathcal{N}}_7(r,s,t) - \frac{2^3}{3^3}b_{10}(r,s,t)$$

$$\mathcal{N}_8(r,s,t) = \tilde{\mathcal{N}}_8(r,s,t) - \frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t) - \frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)$$

$$\mathcal{N}_9(r,s,t) = \frac{64}{63}b_9(r,s,t) - \frac{8}{63}b_{10}(r,s,t)$$

$$\mathcal{N}_{10}(r,s,t) = -\frac{8}{63}b_9(r,s,t) + \frac{64}{63}b_{10}(r,s,t)$$

These are somewhat complex forms for the basis functions so we wish to verify the simple property $\sum \mathcal{N}_i(r,s,t) = 1$ for all (r,s,t) inside the element:

$$
\begin{aligned}
\sum_{i=1}^{10} \mathcal{N}_i(r,s,t) &= \sum_{i=1}^{8} \tilde{\mathcal{N}}_i(r,s,t) \\
&+ \left[\frac{2^3}{3^3}\left(-1 - \frac{10}{21} - \frac{4}{21} - \frac{10}{21} - \frac{10}{21} - \frac{4}{21} - \frac{4}{21}\right) + \frac{64}{63} - \frac{8}{63}\right] b_9(r,s,t) \\
&+ \left[\frac{2^3}{3^3}\left(-\frac{4}{21} - \frac{10}{21} - \frac{4}{21} - \frac{4}{21} - \frac{10}{21} - 1 - \frac{10}{21}\right) - \frac{8}{63} + \frac{64}{63}\right] b_{10}(r,s,t) \\
&= 1 + \left[\frac{2^3}{3^3}(-1 - 42/21) + \frac{56}{63}\right] b_9(r,s,t) + \left[\frac{2^3}{3^3}(-42/21 - 1) + \frac{56}{63}\right] b_{10}(r,s,t) \\
&= 1 + \left[\frac{2^3}{3^3}(-3) + \frac{8}{9}\right] b_9(r,s,t) + \left[\frac{2^3}{3^3}(-3) + \frac{8}{9}\right] b_{10}(r,s,t) \\
&= 1 \quad (10.23)
\end{aligned}
$$

Let us move to first order consistency with $f(r) = r$:

$$f^h(r,s,t) = \sum_{i=1}^{10} \mathcal{N}_i(r,s,t)f_i = \sum_{i=1}^{10} \mathcal{N}_i(r,s,t)r_i \quad (10.24)$$

It has been established for the $\tilde{\mathcal{N}}_i$ functions so we are left with

$$
\begin{aligned}
f^h(r,s,t) \;=\; & \underbrace{\sum_{i=1}^{8}\tilde{\mathcal{N}}_i r_i}_{=r} \\
& -\frac{2^3}{3^3}b_9(r,s,t)(-1) \\
& -\frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t)(+1) -\frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)(+1) \\
& -\frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t)(+1) -\frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)(+1) \\
& -\frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t)(-1) -\frac{2^3}{3^3}\frac{4}{21}b_{10}r,s,t)(-1) \\
& -\frac{2^3}{3^3}\frac{10}{21}b_9(r,s,t)(-1) -\frac{2^3}{3^3}\frac{4}{21}b_{10}(r,s,t)(-1) \\
& -\frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t)(+1) -\frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)(+1) \\
& -\frac{2^3}{3^3}b_{10}(r,s,t)(+1) \\
& -\frac{2^3}{3^3}\frac{4}{21}b_9(r,s,t)(-1) -\frac{2^3}{3^3}\frac{10}{21}b_{10}(r,s,t)(-1) \\
& +\frac{64}{63}b_9(r,s,t)(-1/3) -\frac{8}{63}b_{10}(r,s,t)(-1/3) \\
& -\frac{8}{63}b_9(r,s,t)(+1/3) +\frac{64}{63}b_{10}(r,s,t)(+1/3) \\
\;=\; & r + b_9(r,s,t)\left(\frac{8}{27}-\frac{8}{27}\frac{10}{21}-\frac{8}{27}\frac{4}{21}+\frac{8}{27}\frac{10}{21}+\frac{8}{27}\frac{10}{21}-\frac{8}{27}\frac{4}{21}+\frac{8}{27}\frac{4}{21}-\frac{64}{189}-\frac{8}{189}\right) \\
+ \;& b_{10}(r,s,t)\left(-\frac{8}{27}\frac{4}{21}-\frac{8}{27}\frac{10}{21}+\frac{8}{27}\frac{4}{21}+\frac{8}{27}\frac{4}{21}-\frac{8}{27}\frac{10}{21}-\frac{8}{27}+\frac{8}{27}\frac{10}{21}+\frac{8}{189}+\frac{64}{189}\right) \\
\;=\; & r + b_9(r,s,t)\frac{8}{27}\underbrace{\left(1-\frac{10}{21}-\frac{4}{21}+\frac{10}{21}+\frac{10}{21}-\frac{9}{7}\right)}_{=0} +b_{10}(r,s,t)\frac{8}{27}\underbrace{\left(-\frac{4}{21}-\frac{10}{21}+\frac{4}{21}+\frac{4}{21}-\frac{10}{21}-\right.}_{=0} \\
\;=\; & r
\end{aligned}
$$

which proves first-order consistency.

The derivatives of the $\tilde{\mathcal{N}}_i$ basis functions are already established so we only focus on the spatial

derivatives of the bubble functions:

$$
\begin{aligned}
\frac{\partial b_9}{\partial r} &= \left(\frac{27}{32}\right)^3 (1-s^2)(1-t^2)(1-s)(1-t)(-1-2r+3r^2) \\
\frac{\partial b_9}{\partial s} &= \left(\frac{27}{32}\right)^3 (1-r^2)(1-t^2)(1-r)(1-t)(-1-2s+3s^2) \\
\frac{\partial b_9}{\partial t} &= \left(\frac{27}{32}\right)^3 (1-r^2)(1-s^2)(1-r)(1-s)(-1-2t+3t^2) \\
\frac{\partial b_{10}}{\partial r} &= \left(\frac{27}{32}\right)^3 (1-s^2)(1-t^2)(1+s)(1+t)(1-2r-3r^2) \\
\frac{\partial b_{10}}{\partial s} &= \left(\frac{27}{32}\right)^3 (1-r^2)(1-t^2)(1+r)(1+t)(1-2s-3s^2) \\
\frac{\partial b_{10}}{\partial t} &= \left(\frac{27}{32}\right)^3 (1-r^2)(1-s^2)(1+r)(1+s)(1-2t-3t^2)
\end{aligned}
$$

🔖 Relevant Literature: Fortin and Fortin [429] (1985), Soulaimani, Fortin, Ouellet, Dhatt, and Bertrand [1261] (1987)

Bishnu talksabout Nitsche bc ? press error near boundary in fofo85

10.4.18 The DSSY pair in 3D

What follows is mostly from From Jang *et al.* (2005) [667].

The non-conforming finite element space is defined based on the reference cubic element Q_l on $[-1,1]^3$:

$$
Q_l = \text{Span}\{1, r, s, t, \theta_l(r) - \theta_l(s), \theta_l(r) - \theta_l(t)\} \qquad l = 1, 2
$$

with[12]

$$
\begin{aligned}
\theta_1(r) &= r^2 - \frac{5}{3}r^4 \\
\theta_2(r) &= r^2 - \frac{25}{6}r^4 + \frac{7}{2}r^6
\end{aligned}
\tag{10.26}
$$

The dimension of Q_l is six and the θ_l functions are as follows:
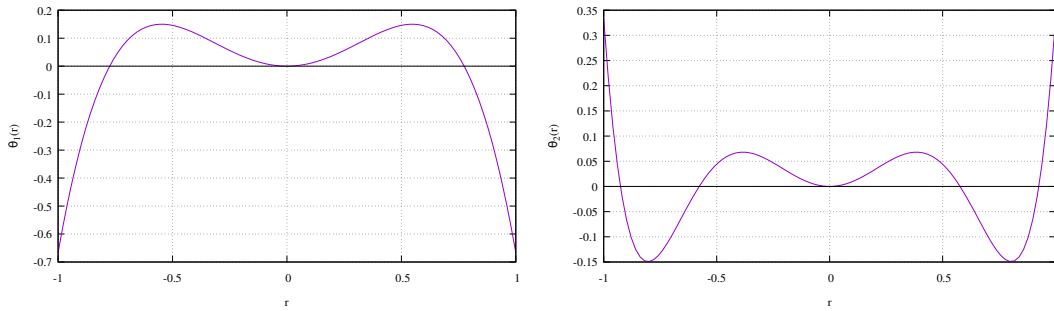
Representation of functions θ_1 (left) and θ_2 (right).

We have:

- $\theta_1(r=-1) = \theta_1(r=+1) = -\frac{2}{3}$, $\theta_1(r=0) = 0$

- $\theta_2(r=-1) = \theta_2(r=+1) = \frac{1}{3}$, $\theta_2(r=0) = 0$

---

[12]Douglas *et al.* [367], Eq. 2.20

These functions have the property that their average on an edge is zero:

$$
\frac{1}{1-(-1)} \int_{-1}^{+1} \theta_1(r)dr = \frac{1}{2} \int_{-1}^{+1} \left(r^2 - \frac{5}{3}r^4 \right) dr
$$
$$
= \frac{1}{2} \left(\frac{2}{3} - \frac{5}{3}\frac{2}{5} \right)
$$
$$
= 0
$$
$$
\frac{1}{1-(-1)} \int_{-1}^{+1} \theta_2(r)dr = \frac{1}{2} \int_{-1}^{+1} \left(r^2 - \frac{25}{6}r^4 + \frac{7}{2}r^6 \right) dr
$$
$$
= \frac{1}{2} \left(\frac{2}{3} - \frac{25}{6}\frac{2}{5} + \frac{7}{2}\frac{2}{7} \right) dr
$$
$$
= 0
$$

The element has 6 nodes that are located at the face centers of a cube or a brick. The pointwise continuity between interfacing elements is guaranteed only at the face centers, so the field quantities are not conforming along the interface [667].
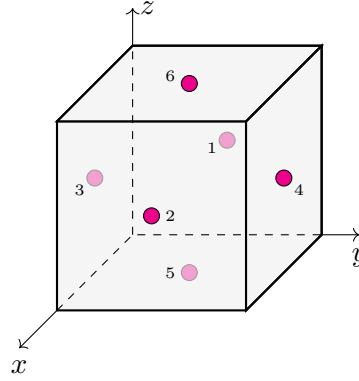
(axis should actually be in the middle of the cube!)

The six nodes of the reference element are

$\vec{r}_1$	-1	0	0
$\vec{r}_2$	$+1$	0	0
$\vec{r}_3$	0	-1	0
$\vec{r}_4$	0	$+$	0
$\vec{r}_5$	0	0	-1
$\vec{r}_6$	0	0	$+1$

and their corresponding basis functions:

$$
N_1^{(l)}(r,s,t) = \frac{1}{6} - \frac{1}{2}r + \frac{1}{6\theta_l(1)}(2\theta_l(r) - \theta_l(s) - \theta_l(t))
$$
$$
N_2^{(l)}(r,s,t) = \frac{1}{6} + \frac{1}{2}r + \frac{1}{6\theta_l(1)}(2\theta_l(r) - \theta_l(s) - \theta_l(t))
$$
$$
N_3^{(l)}(r,s,t) = \frac{1}{6} - \frac{1}{2}s + \frac{1}{6\theta_l(1)}(2\theta_l(s) - \theta_l(t) - \theta_l(r))
$$
$$
N_4^{(l)}(r,s,t) = \frac{1}{6} + \frac{1}{2}s + \frac{1}{6\theta_l(1)}(2\theta_l(s) - \theta_l(t) - \theta_l(r))
$$
$$
N_5^{(l)}(r,s,t) = \frac{1}{6} - \frac{1}{2}t + \frac{1}{6\theta_l(1)}(2\theta_l(t) - \theta_l(r) - \theta_l(s))
$$
$$
N_6^{(l)}(r,s,t) = \frac{1}{6} + \frac{1}{2}t + \frac{1}{6\theta_l(1)}(2\theta_l(t) - \theta_l(r) - \theta_l(s)) \tag{10.27}
$$

These basis functions are also in [367]. We can easily verify that $\sum_i N_i(r, s, t) = 1$ and that $N_i(\vec{r}_j) = \delta_{ij}$:

$$N_1^{(l)}(r_1, s_1, t_1) = \frac{1}{6} - \frac{1}{2}(-1) + \frac{1}{6\theta_l(1)}(2\theta_l(-1) - \theta_l(0) - \theta_l(0)) = 1$$

$$N_1^{(l)}(r_2, s_2, t_2) = \frac{1}{6} - \frac{1}{2}(+1) + \frac{1}{6\theta_l(1)}(2\theta_l(+1) - \theta_l(0) - \theta_l(0)) = 0$$

$$N_1^{(l)}(r_3, s_3, t_3) = \frac{1}{6} - \frac{1}{2}(0) + \frac{1}{6\theta_l(1)}(2\theta_l(0) - \theta_l(-1) - \theta_l(0)) = 0$$

$$N_1^{(l)}(r_4, s_4, t_4) = \frac{1}{6} - \frac{1}{2}(0) + \frac{1}{6\theta_l(1)}(2\theta_l(0) - \theta_l(+1) - \theta_l(0)) = 0$$

$$N_1^{(l)}(r_5, s_5, t_5) = \frac{1}{6} - \frac{1}{2}(0) + \frac{1}{6\theta_l(1)}(2\theta_l(0) - \theta_l(0) - \theta_l(-1)) = 0$$

$$N_1^{(l)}(r_6, s_6, t_6) = \frac{1}{6} - \frac{1}{2}(0) + \frac{1}{6\theta_l(1)}(2\theta_l(0) - \theta_l(0) - \theta_l(+1)) = 0$$

etc ...

$$\partial_r N_1^{(l)}(r, s, t) = -\frac{1}{2} + \frac{1}{3\theta_l(1)}\theta_l'(r)$$

$$\partial_r N_2^{(l)}(r, s, t) = +\frac{1}{2} + \frac{1}{3\theta_l(1)}\theta_l'(r)$$

$$\partial_r N_3^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(r)$$

$$\partial_r N_4^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(r)$$

$$\partial_r N_5^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(r)$$

$$\partial_r N_6^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(r) \tag{10.28}$$

$$\partial_s N_1^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(s)$$

$$\partial_s N_2^{(l)}(r, s, t) = -\frac{1}{6\theta_l(1)}\theta_l'(s)$$

$$\partial_s N_3^{(l)}(r, s, t) = -\frac{1}{2} + \frac{1}{3\theta_l(1)}\theta_l'(s)$$

$$\partial_s N_4^{(l)}(r, s, t) = +\frac{1}{2} + \frac{1}{3\theta_l(1)}\theta_l'(s)$$

$$\partial_s N_5^{(l)}(r, s, t) = \frac{1}{6\theta_l(1)}\theta_l'(s)$$

$$\partial_s N_6^{(l)}(r, s, t) = \frac{1}{6\theta_l(1)}\theta_l'(s) \tag{10.29}$$

$$\partial_t N_1^{(l)}(r,s,t) = -\frac{1}{6\theta_l(1)}\theta_l(t)$$

$$\partial_t N_2^{(l)}(r,s,t) = -\frac{1}{6\theta_l(1)}\theta_l(t)$$

$$\partial_t N_3^{(l)}(r,s,t) = -\frac{1}{6\theta_l(1)}\theta_l(t)$$

$$\partial_t N_4^{(l)}(r,s,t) = -\frac{1}{6\theta_l(1)}\theta_l(t)$$

$$\partial_t N_5^{(l)}(r,s,t) = -\frac{1}{2}t + \frac{1}{3\theta_l(1)}\theta_l(t)$$

$$\partial_t N_6^{(l)}(r,s,t) = +\frac{1}{2}t + \frac{1}{3\theta_l(1)}\theta_l(t) \tag{10.30}$$

🔷 Relevant Literature: Douglas *et al.* (1999) [367],

10.4.19 Other FE element pairs

- $\boldsymbol{Q}_2 \times Q_2$: This element is never used, probably because a) it is unstable, b) it is very costly. There is one reference to it in Hughes, Franca, and Balestra [639] (1986).

- $\boldsymbol{Q}_1 \times P_{-1}$ Bilinear velocities, piecewise linear discontinuous polynomial pressure.

- See Fortin [427] for various stable low order elements other than the enriched $\boldsymbol{Q}_1^+ \times P_0$

- $\boldsymbol{Q}_1 \times Q_1$ + nonconforming null edge average [437]

10.5 Which velocity-pressure pair for Stokes? (triangles & tetra)

10.5.1 The $P_1 \times P_0$ pair

The standard form of the $\boldsymbol{P}_1 \times P_0$ element is as follows:

```
*
|\   P1      |\    P0
| \          | \
|  \         |  \
|   \        |   \
|    \       |    \
|     \      | o   \
|      \     |      \
*-------*    ---------
```
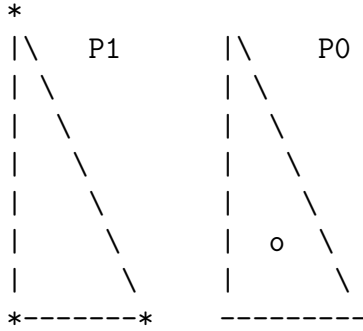
The pressure is constant inside the element and discontinuous from element to element.

Qin and Zhang [1092] (2007) state:

> The element is unstable for any mesh since the dimension of the discrete velocity space is always less than that of the pressure space (with Dirichlet boundary condition). However, this element provides optimal approximations for both the velocity and the pressure on many mesh families.
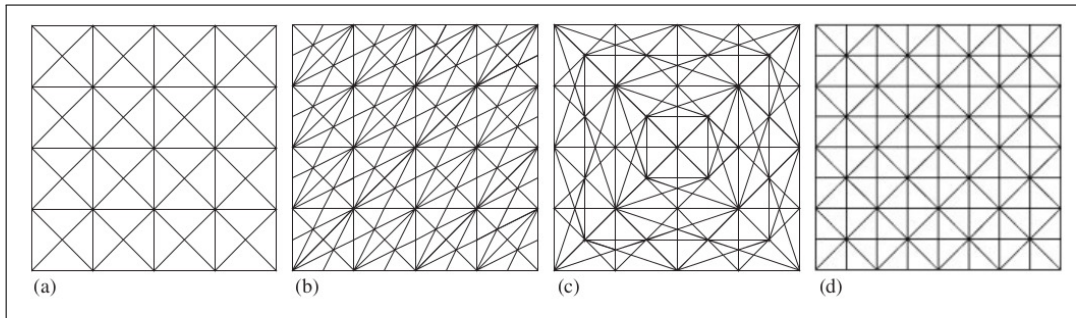
A crisscross, a Powell–Sabin, a Powell–Sabin–Heindl and a type-2 grid. Taken from [1091].

Arnold [28] (1993) states:

> Unfortunately, this simplest possible Stokes element is notoriously unstable. On any triangulation with at least three vertices on the boundary the dimension of the pressure space exceeds that of the velocity space [...] and the finite dimensional problem is singular. Moreover, while the discrete velocity field u_h is uniquely determined (as it is for any conforming method for the Stokes problem), for this choice of elements u_h belongs to the space of divergence-free fields piecewise linear fields, and on many meshes, for example on a uniform diagonal mesh of the square [...], this space is known to reduce to zero. So even after accounting for the indeterminancy of the pressure we have no convergence.

Example 3.2 of Boffi, Brezzi, and Fortin [114] (2008) explain neatly the locking phenomenon and how to circumvent it via a so-called cross-grid macroelement. See also Hong, Kim, and Lee [621] (2003).

Note that Zhang [1512] (2008) establishes the stability of the $\boldsymbol{P}_1 \times P_0$ mixed-element on general Powell-Sabin triangular grids.

The piecewise linear finite element solution approximating the velocity is divergence-free pointwise for the Stokes equations. The finite element solution approximating the pressure in the Stokes equations can be obtained as a byproduct if an iterative method is adopted for solving the discrete linear system of equations. Numerical tests are presented confirming the theory on the stability and the optimal order of convergence for the P_1 Powell-Sabin divergence-free finite element method.

Fig. 3.2. A uniform triangulation $\tilde{\Omega}_h$ splits into a Powell-Sabin grid Ω_h.
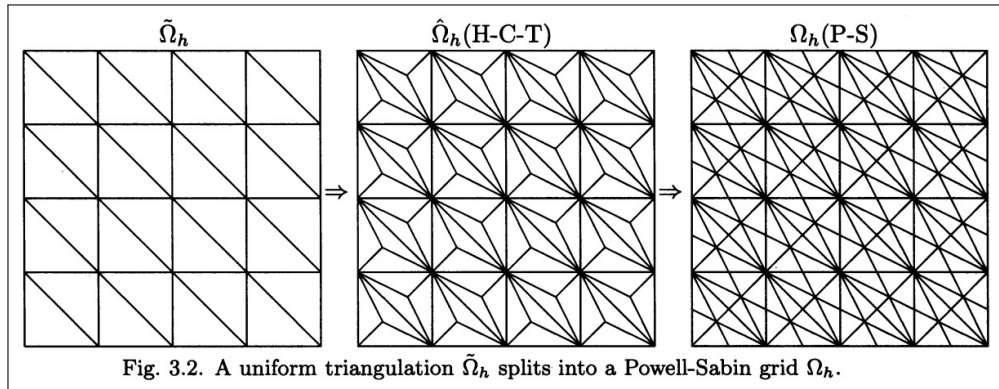
Taken from Zhang [1512], 2008. Looking at this node layout we see that this element pair will not be trivial to implement since the velocity nodes inside the element are placed like for the P_2^+ element but instead make up 6 smaller elements.

C_0–P_1 velocity on Ω_h, C_{-1}–P_0 pressure on $\tilde{\Omega}_h$

Fig. 3.4. A stable pair of mixed finite elements (Powell-Sabin grids for the velocity).
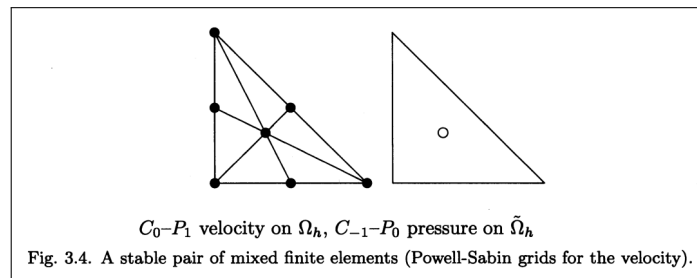
Taken from Zhang [1512], 2008.

Note that the figure above is *very* important: the 'master' triangle is divided into 6 smaller P_1 triangles for the velocity but the pressure is P_0 on the master triangle, not on the 6 smaller ones!

In his lecture notes[13], Guermond states

> A simple alternative to the $Q_1 \times P_0$ element consists of using the $P_1 \times P_0$ element. Let $\mathcal{T}_h$ be a mesh of D composed of affine simplices, and approximate the velocity with continuous piecewise linear polynomials and the pressure with (discontinuous) piecewise constants. Since the velocity is piecewise linear, its divergence is constant on each simplex. As a result, testing the divergence of the velocity with piecewise constants enforces the divergence to be zero everywhere. That is to say, the $P_1 \times P_0$ finite element yields a velocity approximation that is exactly divergence-free [...]. Unfortunately, this pair does not satisfy the inf-sup condition.

I have not found a paper yet which showcases its accuracy on a manufactured solution and compares it to other element pairs.

---

[13]https://www.math.tamu.edu/~guermond/
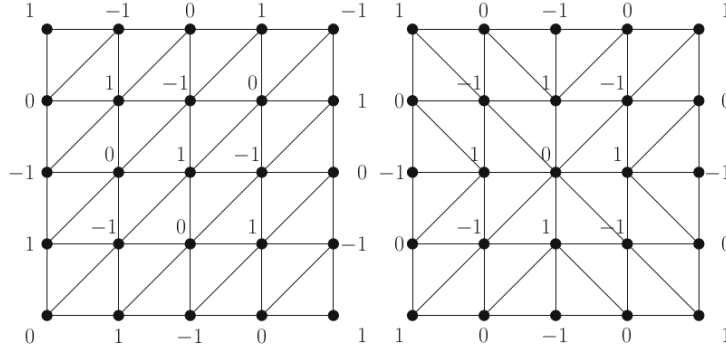
Taken from John [688], page 65. On these meshes, P_1 possesses 25 degrees of freedom (number of vertices) whereas P_0 has even 32 degrees of freedom (number of mesh cells). There is one degree of freedom less in the corresponding pressure finite element spaces since the discrete pressure has to fulfill the integral mean value condition. Thus, spurious modes have to be expected.

Elman Silvester Wathen say (5.3.3) that

> It can be readily stabilized using the pressure jump stabilization together with an appropriate macroelement subdivision.

See Norburn and Silvester [1002] (1998) for globally and locally stabilised versions.

In Carey and Oden [224] (1986), we read[14]:

> Another interesting choice is the triangle with linear velocity and piecewise-constant pressure. The vertices of the triangle are velocity nodes and the element centroids may be taken as pressure nodes. This element element would appear to be the simplest and most natural choice for a mixed method, but again the consistency condition fails. An easy physical interpretation is possible in this case: As $h \to 0$ the number of triangles increases faster than the number of vertices. Concequently, the dimension of the pressure space P^h grows to exceed that of the velocity space V^h. For example, consider the unit square with an $N \times N$ rectangular mesh subdivided into $2N^2$ triangles. We have $2N^2$ pressure nodes (and basis functions) but only $2(N-1)^2$ interior velocity degrees of freedom (and basis functions) if Dirichlet data is specified. Each pressure equation is a discrete constraint equation, and the problem is overconstrained. The velocity $\vec{v}_h = \vec{0}$, which satisfies the mass constraint but not the momentum equations.
>
> [..]
>
> A balance between the dimensions of the velocity and pressure spaces should be maintained as the mesh size h decreases. That is, if the dimension of the pressure space P^h is too large relative to that of the velocity space as the mesh is refined, the discrete constraint equations dominate the discrete momentum equations. It has been argued from this standpoint that elements should be chosen such that $dim\, P^h < dim\, V^h$. While the underlying reasoning is plausible and this gives a good guideline to follow in practice, it is neither necessary nor sufficient. In fact the inf-sup condition provides a mathematical theory for selecting elements.

10.5.2 The $P_1^+ \times P_0$ pair

---

[14]page 138

This element pair counts $8 * 3 = 24$ vdofs and 1 pdof.

In Table I of Bertrand, Gadbois, and Tanguy [89] (1992) we find:

$$
\begin{aligned}
\mathcal{N}_1(r,s,t) &= (1 - r - s - t) - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7) & \vec{r}_1 &= (0,0,0) \\
\mathcal{N}_2(r,s,t) &= r - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_8) & \vec{r}_2 &= (1,0,0) \\
\mathcal{N}_3(r,s,t) &= s - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}_8) & \vec{r}_3 &= (0,1,0) \\
\mathcal{N}_4(r,s,t) &= t - \frac{1}{3}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8) & \vec{r}_4 &= (0,0,1) \\
\mathcal{N}_5(r,s,t) &= 27(1 - r - s - t)rs & \vec{r}_5 &= (\frac{1}{3}, \frac{1}{3}, 0) \\
\mathcal{N}_6(r,s,t) &= 27(1 - r - s - t)rt & \vec{r}_6 &= (\frac{1}{3}, 0, \frac{1}{3}) \\
\mathcal{N}_7(r,s,t) &= 27(1 - r - s - t)st & \vec{r}_7 &= (0, \frac{1}{3}, \frac{1}{3}) \\
\mathcal{N}_8(r,s,t) &= 27rst & \vec{r}_8 &= (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})
\end{aligned}
\tag{10.31}
$$

Note that we can verify that

$$
\begin{aligned}
\sum_{i=1}^{8} \mathcal{N}_i(r,s,t) &= \mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3 + \mathcal{N}_4 + \mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8 \\
&= (1 - r - s - t) - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}7) \\
&+ r - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}8) + s - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}8) + t - \frac{1}{3}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}8) \\
&+ \mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8 \\
&= 1 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3} + 1)\mathcal{N}_5 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3} + 1)\mathcal{N}_6 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3} + 1)\mathcal{N}_7 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3} \\
&= 1
\end{aligned}
$$

The authors report a 1st order convergence for the velocity on a manufactured solution. The authors state:

> The shape functions of the original P_1 basis do not vanish at the newly created mid-face nodes. This means that the additional degrees of freedom at these nodes do not represent velocity as such, but rather velocity corrections and must be treated accordingly.

10.5.3 The $P_1 \times P_0$ pair of Crouzeix-Raviart

This is the degree-1 Crouzeix–Raviart on a triangle. See Crouzeix and Raviart [309] (1973).
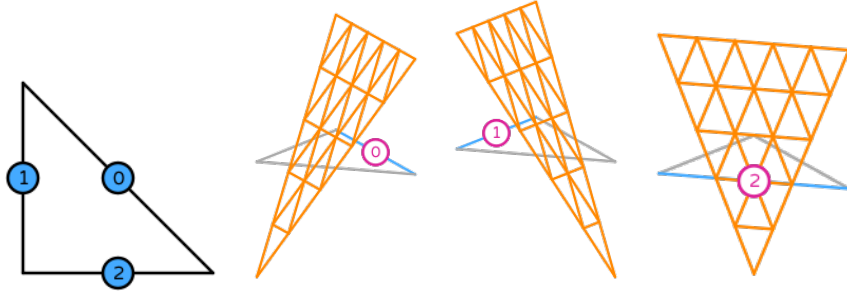
We find on the defelement page[15] that the basis functions are given by:

$$\begin{aligned}
\mathcal{N}_0(r,s) &= 2r + 2s - 1 \\
\mathcal{N}_1(r,s) &= 1 - 2r \\
\mathcal{N}_2(r,s) &= 1 - 2s
\end{aligned}$$

We find $\mathcal{N}_i(r_j, s_j) = \delta_{ij}$.

10.5.4 The $P_1^{++} \times P_{-1}$ pair

<div align="right"><code>pair_p1ppp1.tex</code></div>

Taken from Bertrand, Gadbois, and Tanguy [89] (1992). dots, crosses and Y are all velocity nodes. The black square stands for a discontinous linear pressure.

This element pair counts $(4 + 4 + 1) * 3 = 27$ vdofs and 3 pdof.
In Table I of Bertrand, Gadbois, and Tanguy [89] (1992) we find:

$$\begin{aligned}
\mathcal{N}_1(r,s,t) &= (1 - r - s - t) - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7) - \frac{1}{4}\mathcal{N}_9 & \vec{r}_1 &= (0,0,0) \\
\mathcal{N}_2(r,s,t) &= r - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 & \vec{r}_2 &= (1,0,0) \\
\mathcal{N}_3(r,s,t) &= s - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 & \vec{r}_3 &= (0,1,0) \\
\mathcal{N}_4(r,s,t) &= t - \frac{1}{3}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 & \vec{r}_4 &= (0,0,1) \\
\mathcal{N}_5(r,s,t) &= 27(1 - r - s - t)rs - \frac{108}{256}\mathcal{N}_9 & \vec{r}_5 &= (\tfrac{1}{3}, \tfrac{1}{3}, 0) \\
\mathcal{N}_6(r,s,t) &= 27(1 - r - s - t)rt - \frac{108}{256}\mathcal{N}_9 & \vec{r}_6 &= (\tfrac{1}{3}, 0, \tfrac{1}{3}) \\
\mathcal{N}_7(r,s,t) &= 27(1 - r - s - t)st - \frac{108}{256}\mathcal{N}_9 & \vec{r}_7 &= (0, \tfrac{1}{3}, \tfrac{1}{3}) \\
\mathcal{N}_8(r,s,t) &= 27rst - \frac{108}{256}\mathcal{N}_9 & \vec{r}_8 &= (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}) \\
\mathcal{N}_9(r,s,t) &= 256(1 - r - s - t)rst & \vec{r}_9 &= (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4})
\end{aligned}$$

---

[15]`https://defelement.com/elements/examples/triangle-crouzeix-raviart-1.html`

Note that we can verify that

$$
\begin{aligned}
\sum_{i=1}^{8} \mathcal{N}_i(r,s,t) &= \mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3 + \mathcal{N}_4 + \mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8 + \mathcal{N}_9 \\
&= (1 - r - s - t) - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7) - \frac{1}{4}\mathcal{N}_9 \\
&+ \; r - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 \\
&+ \; s - \frac{1}{3}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 \\
&+ \; t - \frac{1}{3}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{4}\mathcal{N}_9 \\
&+ \; +\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_8 + \mathcal{N}_9 \\
&= 1 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3}+)\mathcal{N}_5 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3}+)\mathcal{N}_6 \\
&+ (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3}+)\mathcal{N}_7 + (-\frac{1}{3} - \frac{1}{3} - \frac{1}{3}+)\mathcal{N}_8 + (-\frac{1}{4} - \frac{1}{4} - \frac{1}{4} - \frac{1}{4} + 1)\mathcal{N}_9 \\
&= 1
\end{aligned}
\tag{10.33}
$$

The authors report a 1st order convergence for the velocity on a manufactured solution.

The authors state:

> The shape functions of the original P_1 basis do not vanish at the newly created mid-face nodes. This means that the additional degrees of freedom at these nodes do not represent velocity as such, but rather velocity corrections and must be treated accordingly.

10.5.5 The $P_2 \times P_0$ pair

pair_p2p0.tex

[114], [**cakp18**] stable (Kanschat book)

compared with P1NC-P0 and BR element in Carstensen, Köhler, Peterseim, and Schedensack [228] (2015).

10.5.6 The $P_1 \times P_1$-stabilised pair

pair_p1p1stab.tex

Like its quadrilateral counterpart $\boldsymbol{Q}_1 \times Q_1$, the $P_1 \times P_1$ pair is not stable and needs to be stabilised [1002, 1308]. TerraNeo code uses stabilised with PSPG [61].

Norburn and Silvester [1002]

10.5.7 The $P_1^+ \times P_1$ (MINI) pair in 2D & 3D

pair_mini.tex

The MINI element was first introduced in Arnold, Brezzi, and Fortin [26] (1984). It is also denoted by $\boldsymbol{P}_1^+ \times P_1$. It is also discussed in Section 3.6.1 of John [688] (2016) and in Section 6.1 of Boffi, Brezzi, and Fortin [114] (2008). It is thoroughly studied in Cioncolini and Boffi [276] (2019).

As explained in Braess [136], since the support of the cubic bubble is restricted to the element, the associated variable (dofs living on the bubble) can be eliminated from the resulting system of linear equations by static condensation. Also, the MINI element is cheaper than the Taylor-Hood element but it is commonly accepted that it yields a poorer approximation of the pressure.

4 vel. nodes, 3 press. nodes

Remark. *Note that Franca and Oliveira [436] (2003) propose an equal-order-linear-continuous velocity-pressure variables which is enriched with velocity and pressure bubble functions to model the Stokes problem. They show by static condensation that these bubble functions give rise to a stabilized method involving least-squares forms of the momentum and of the continuity equations. In some cases their approach recovers the MINI element. Also check Ganesan, Matthies, and Tobiska [460] (2008).*

The 3D MINI element is not very common but it is used for instance in Pichelin and Coupez [1059] (1998) or Tommasi, Knoll, Vauchez, Signorelli, Thoraval, and Logé [1360] (2009). It is also said to be LBB stable in Reddy [1118, p180]. It is used in [453] phd thesis in the context of microstructures deformation modeling, which itself cites Cao, Montmitonnet, and Bouchard [222] (2013).
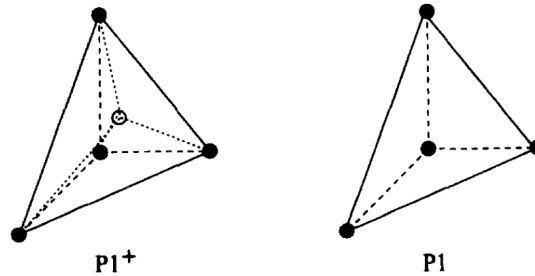
Velocity and pressure nodes for the 3D MINI element, taken from [1059]

Note that this element is used in Braess & Wriggers (2000) [138] in the context of Arbitrary Lagrangian Eulerian finite element analysis of free surface flows, and also in Zlotnik, Diez, Fernandez, and Verges [1547] (2007) for subduction with X-FEM technique. . It is also mentioned in Nafa and Thatcher [974] (1993). The 2D element is implemented in STONE 120.

https://defelement.com/elements/mini.html

https://www.math.uci.edu/~chenlong/ifemdoc/fem/StokesP1bP1femrate.html

10.5.8   The $P_2 \times P_1$ pair

From Segal [1224]:

> Taylor-Hood elements [1326] are character-
> ized by the fact that the pressure is contin-
> uous in the region Ω. A typical example is
> the quadratic triangle ($\boldsymbol{P}_2 \times P_1$ element). In
> this element the velocity is approximated
> by a quadratic polynomial and the pressure
> by a linear polynomial. One can easily ver-
> ify that both approximations are continu-
> ous over the element boundaries.

6 vel. nodes, 3 press. nodes

$\bullet\,\vec{v}$ $\bigcirc\,p$

It can be shown, Segal (1979), that this element is ad-
missible if at least 3 elements are used. The quadrilat-
eral counterpart of this triangle is the $\boldsymbol{Q}_2 \times Q_1$ element.
Reddy and Gartling [1118, p179] also report this ele-
ment to be LBB stable. It is also mentioned in Nafa
and Thatcher [974].

📚 Relevant Literature: Schubert and Anderson [1218]
(1985), Leng, Ju, Xie, Cui, and Gunzburger [817]
(2014), Cuffaro, Miglio, Penati, and Viganò [312]
(2020)

Note that the element is not stable if the triangulation contains less than three triangles (Lemma
3.125 of [688] on page 100).

Optimal error estimates are derived for this element pair in Stenberg [1290] (1987).

10.5.9   The $P_2^+ \times P_{-1}$ pair (Crouzeix-Raviart)

Since the $\boldsymbol{P}_2 \times P_{-1}$ pair is not LBB stable [1118, p179], (see also table 3.13-1 of Gresho and Sani
[517]) it is enhanced by a cubic bubble and is therefore called $\boldsymbol{P}_2^+ \times P_{-1}$.

This element was first introduced in Crouzeix and Raviart [309] (1973). It is a seven-node
triangle (in 2D) with quadratic velocity shape functions enhanced by a cubic bubble function and
discontinuous linear interpolation for the pressure field [317]. This element is LBB stable and no
additional stabilization techniques are required [394]. The '+' in its name stands for the bubble
while the '-' stands for the discontinuous character of the pressure field: it is P_1 over the element,
but discontinuous across element edges.

Remark. *Cuvelier et al. , 1986 [317] recommend a 6-point or 7-point quadrature rule for this element.*

Remark. *Segal [1224] explains for output purposes (printing, plotting etc.) the discontinuous pres-
sures are averaged in vertices for all the adjoining elements. See also Fig. 7.3 of [317].*

Remark. *The simplest Crouzeix-Raviart element is the non-conforming linear triangle with constant
pressure [317], see Section 10.5.17.*

It is worth noting that this element has more degrees of freedom than the Taylor-Hood element for
the same order of accuracy. However, since the bubble can be eliminated, one can design a modified
version of this element.

Check Cuvelier book chapter 8 for modified element
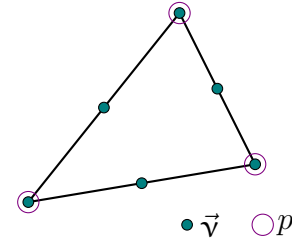
Remark. *I have once asked the (main) author of MILAMIN why he chose this element, for example over the $\boldsymbol{P}_2 \times P_1$. His answer is as follows: "Elements with continuous pressure are incapable of converging in the Linf norm for mechanical problems exhibiting pressure jumps such as the inclusion-host setup. During my MSc and PhD I was focusing on sharp heterogeneities, so this is why I decided to choose $\boldsymbol{P}_2^+ \times P_{-1}$. You will see that it is also easy to invert the pressure mass matrix for such elements, which is really useful (both for the augmentation and preconditioning)."*

This element is used by Poliakov and Podlachikov [1073] (1992) to study the deformation of the surface above a rising diapir. Note that they actually use a "13 point integration formula (Hughes 1987) for calculation of the stiffness matrix was used in order to conserve detailed information from the marker field in the coarse FEM mesh". It is the element used in the MILAMIN code, see Dabrowski, Krotkiewski, and Schmid [318] (2008). It is also featured in Thieulot and Bangerth [1346] (2025). It is also used in Andrés-Martínez, Morgan, Pérez-Gussinyé, and Rüpke [24] (2015) in the context of a new free-surface stabilization scheme. It is the element used in LaCoDe, see de Montserrat, Morgan, and Hasenclever [341] (2019). It is mentioned in Section 6.2 in Boffi, Brezzi, and Fortin [114] (2008). It is compared to the $\boldsymbol{P}_2 \times P_1$ element for the Navier-Stokes equations in Krahl and Bänsch [773] (2005).

Taken from Bertrand, Gadbois, and Tanguy [89] (1992).

This element counts $(4 + 6 + 4 + 1) * 3 = 45$ vdofs and 4 pdofs.

The 3D version of this element is presented in Bertrand, Gadbois, and Tanguy [89] (1992). In their Table III we find the following node coordinates and basis functions:

# node	coordinates	basis function
1	$(0,0,0)$	$\mathcal{N}_1(r,s,t) = (1-r-s-t) - \frac{1}{2}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{12} + \mathcal{N}_{13}) - \mathcal{N}_{15}/4$
2	$(1,0,0)$	$\mathcal{N}_2(r,s,t) = r - \frac{1}{2}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_9) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{12} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
3	$(0,1,0)$	$\mathcal{N}_3(r,s,t) = s - \frac{1}{2}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_{10}) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{13} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
4	$(0,0,1)$	$\mathcal{N}_4(r,s,t) = t - \frac{1}{2}(\mathcal{N}_8 + \mathcal{N}_9 + \mathcal{N}_{10}) - \frac{1}{3}(\mathcal{N}_{12} + \mathcal{N}_{13} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
5	$(\frac{1}{2},0,0)$	$\mathcal{N}_5(r,s,t) = 4(1-r-s-t)r - \frac{4}{9}(\mathcal{N}_{11} + \mathcal{N}_{12}) - \mathcal{N}_{15}/4$
6	$(\frac{1}{2},\frac{1}{2},0)$	$\mathcal{N}_6(r,s,t) = 4rs - \frac{4}{9}(\mathcal{N}_{11} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
7	$(0,\frac{1}{2},0)$	$\mathcal{N}_7(r,s,t) = 4(1-r-s-t)s - \frac{4}{9}(\mathcal{N}_{11} + \mathcal{N}_{13}) - \mathcal{N}_{15}/4$
8	$(0,0,\frac{1}{2})$	$\mathcal{N}_8(r,s,t) = 4(1-r-s-t)t - \frac{4}{9}(\mathcal{N}_{12} + \mathcal{N}_{13}) - \mathcal{N}_{15}/4$
9	$(\frac{1}{2},0,\frac{1}{2})$	$\mathcal{N}_9(r,s,t) = 4rt - \frac{4}{9}(\mathcal{N}_{12} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
10	$(0,\frac{1}{2},\frac{1}{2})$	$\mathcal{N}_{10}(r,s,t) = 4st - \frac{4}{9}(\mathcal{N}_{13} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4$
11	$(\frac{1}{3},\frac{1}{3},0)$	$\mathcal{N}_{11}(r,s,t) = 27(1-r-s-t)rs - \frac{108}{256}\mathcal{N}_{15}$
12	$(\frac{1}{3},0,\frac{1}{3})$	$\mathcal{N}_{12}(r,s,t) = 27(1-r-s-t)rt - \frac{108}{256}\mathcal{N}_{15}$
13	$(0,\frac{1}{3},\frac{1}{3})$	$\mathcal{N}_{13}(r,s,t) = 27(1-r-s-t)st - \frac{108}{256}\mathcal{N}_{15}$
14	$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	$\mathcal{N}_{14}(r,s,t) = 27rst - \frac{108}{256}\mathcal{N}_{15}$
15	$(\frac{1}{4},\frac{1}{4},\frac{1}{4})$	$\mathcal{N}_{15}(r,s,t) = 256(1-r-s-t)rst$

We have

$$
\begin{aligned}
\mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3 + \mathcal{N}_4 \;=\;& (1 - r - s - t) - \frac{1}{2}(\mathcal{N}_5 + \mathcal{N}_7 + \mathcal{N}_8) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{12} + \mathcal{N}_{13}) - \mathcal{N}_{15}/4 \\
+\;& r - \frac{1}{2}(\mathcal{N}_5 + \mathcal{N}_6 + \mathcal{N}_9) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{12} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4 \\
+\;& s - \frac{1}{2}(\mathcal{N}_6 + \mathcal{N}_7 + \mathcal{N}_{10}) - \frac{1}{3}(\mathcal{N}_{11} + \mathcal{N}_{13} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4 \\
+\;& t - \frac{1}{2}(\mathcal{N}_8 + \mathcal{N}_9 + \mathcal{N}_{10}) - \frac{1}{3}(\mathcal{N}_{12} + \mathcal{N}_{13} + \mathcal{N}_{14}) - \mathcal{N}_{15}/4 \\
=\;& 1 - \frac{1}{2}(1+1)\mathcal{N}_5 - \frac{1}{2}(1+1)\mathcal{N}_6 - \frac{1}{2}(1+1)\mathcal{N}_7 - \frac{1}{2}(1+1)\mathcal{N}_8 - \frac{1}{2}(1+1)\mathcal{N}_9 - \frac{1}{2}(1 + 1 \\
& - \frac{1}{3}(1+1+1)\mathcal{N}_{11} - \frac{1}{3}(1+1+1)\mathcal{N}_{12} - \frac{1}{3}(1+1+1)\mathcal{N}_{13} - \frac{1}{3}(1+1+1)\mathcal{N}_{14} - \frac{1}{4}(1 \\
=\;& 1 - \mathcal{N}_5 - \mathcal{N}_6 - \mathcal{N}_7 - \mathcal{N}_8 - \mathcal{N}_9 - \mathcal{N}_{10} - \mathcal{N}_{11} - \mathcal{N}_{12} - \mathcal{N}_{13} - \mathcal{N}_{14} - \mathcal{N}_{15}
\end{aligned}
$$

So in the end

$$
\sum_{i=1}^{15} \mathcal{N}_i = (\mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3 + \mathcal{N}_4) + \sum_{i=5}^{15} \mathcal{N}_i = 1
$$

concerning the 3D version we read in John [688]:

> The enrichment proposed in Crouzeix and Raviart (1973) is given by all mesh cell bubble functions and all face bubble functions. For a mesh cell K with barycentric coordinates λ_1, λ_2, λ_3, λ_4 a mesh cell bubble function is given by $\lambda_1\lambda_2\lambda_3\lambda_4$ and the four face bubble functions are the product of three mutually distinct barycentric coordinates, e.g., $\lambda_1\lambda_2\lambda_3$. Thus, for each velocity component, there are 15 degrees of freedom such that the dimension of the local space for enrichment is 45. From the practical point of view one has to take into consideration that bubbles are higher degree polynomials and their use requires the application of higher order quadrature rules.

> https://defelement.com/elements/conforming-crouzeix-raviart.html

10.5.10   The $P_2^+ \times P_1$ pair

This element pair is not to be mistaken for the Crouzeix-Raviart. Both share the same $\boldsymbol{P}_2^+$ space for the velocity but this element has a continuous linear pressure. It is mentioned in Table 3.13-1 of Gresho and Sani [517]: "LBB stable. Second order. cubic bubble. good element". It is also mentioned in Soulaimani, Fortin, Ouellet, Dhatt, and Bertrand [1261] (1987). It is implemented in STONE   120.

10.5.11   The $P_2 \times (P_1 + P_0)$ pair

This element pair is discussed in 5.3.3 of Elman, Silvester and Wathen:

> [Another] possibility is to construct a hybrid pressure approximation by combining the continuous linear pressure approximation with the discontinuous constant pressure approximation. The resulting mixed method is referred to as the $P_2 - P_{-1*}$ approximation and enjoys the best of both worlds; it has locally incompressibility, and yet it does not have its accuracy compromised by the lower order pressure [of $P_2 \times P_0$]. Perhaps surprisingly, this element is also uniformly stable.
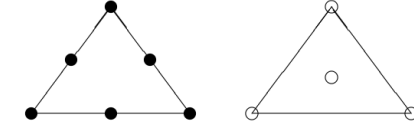
In Gresho & Sani table 3.13-1: "LBB stable yes. Better than P2P1, element mass balance. more work than P2P1. 2 hydrostatic modes. Second order."
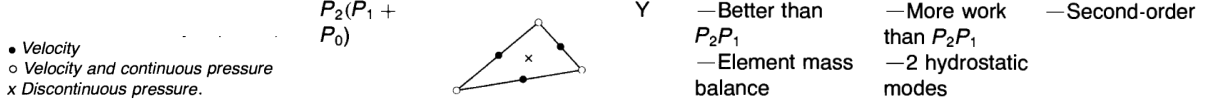
Looking at the figure above, it is clear that the P_1 space is to be understood as a continuous pressure space, with an additional constant bubble.

For the continuous P_1 space, we have the following reference element

$$(r_1, s_1) = (0, 0)$$
$$(r_2, s_2) = (1, 0)$$
$$(r_3, s_3) = (0, 1)$$
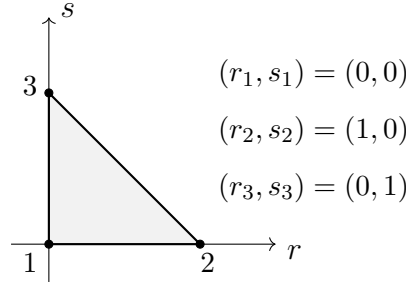
and the basis functions are simply

$$\mathcal{N}_1(r, s) = 1 - r - s \tag{10.38}$$
$$\mathcal{N}_2(r, s) = r \tag{10.39}$$
$$\mathcal{N}_3(r, s) = s \tag{10.40}$$

with the interpolation requirement $\mathcal{N}_i(r_j, s_j) = \delta_{ij}$ fulfilled, as well as $\sum_i \mathcal{N}_i = 1$.

Now, following the figure by Gresho and Sani, I build the reference element for the $P_1 + P_0$ space:

$$(r_1, s_1) = (0, 0)$$
$$(r_2, s_2) = (1, 0)$$
$$(r_3, s_3) = (0, 1)$$
$$(r_4, s_4) = (0.33, 0.33)$$
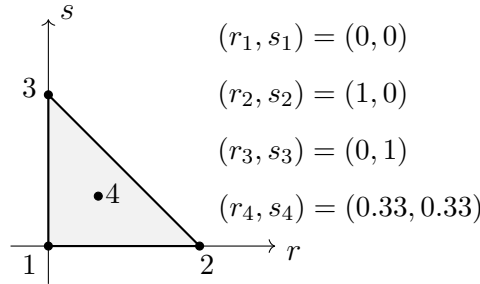
$P_1 + P_0$ means that the pressure inside the element is given by

$$p^h(r, s) = a\mathcal{N}_1(r, s) + b\mathcal{N}_2(r, s) + c\mathcal{N}_3(r, s) + d\mathcal{N}_4(r, s)$$

Note that it is then impossible to find a, b, c, d such that the interpolation requirement $\mathcal{N}_i(r_j, s_j) = \delta_{ij}$ is fulfilled. In other words, the element is not interpolatory, i.e., there is no δ_{ij} property.

With regards to the 'element mass balance', W.B. states :

the mass conservation requires that the function that is constant 1 on one cell and zero on all other cells is part of the function space. That is indeed true – it's the $\mathcal{N}_4$ function. Indeed, that's the purpose of the enrichment with the P_0 part. It is not necessary that *all* shape functions are discontinuous.

Boffi, Cavallini, Gardini, and Gastaldi [116] (2012) state:

> [...] the pressure space Q_h is defined as the sum of two finite element spaces, namely $P_k + P_0$ $(k \geq d - 1)$ [...] for the enhanced Hood–Taylor [...]. However, it can be easily observed that the sum is not direct, since globally constant functions can be represented exactly by means of piecewise P_0 or continuous P_k $(k \geq 1)$ elements. Concerning the implementation of the method, we avoid the computation of the basis functions of such a finite element by testing the discrete problem (2.3) with the basis functions of the two subspaces separately. By the above discussion it turns out that the resulting matrix is rank-deficient, with kernel of dimension 1.

The element pair is also discussed in Chen [239] (2014).

10.5.12 The $P_3 \times P_2$ pair

<div align="right">pair_p3p2.tex</div>

The $\boldsymbol{P}_3 \times P_2$ pair is mentioned in Stenberg [1289] (1990). The P_3 basis functions are presented in Section 8.3.15 and the P_2 basis functions in Section 8.3.13. It is implemented in STONE 120.

10.5.13 The $P_4 \times P_3$ pair

It is implemented in STONE 120. The P_4 and P_3 spaces are defined in Sections 8.3.16 and 8.3.15 respectively.

10.5.14 The Bernardi-Raugel pair

<div align="right">pair_bernardi_raugel.tex</div>

In Carstensen, Köhler, Peterseim, and Schedensack [228] (2015) we find:

> The BR-FEM after Bernardi and Raugel [87] is a modification of the $P_2 \times P_0$ FEM. It is sometimes also called reduced $P_2 \times P_0$ FEM.

They also state that this element also exists in 3D.

(b) BR-FEM
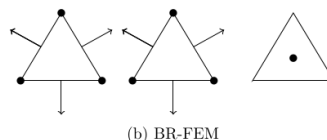
It is also mentioned in Boffi, Brezzi, and Fortin [115] although it seems it is there called the SMALL element (p474).

In Lederer: "Consider the case d = 2. [...] we only need to control the normal velocity at the edge, i.e. adding the edge bubble for both components of the velocity seems to be sub optimal (with respect to computational costs and the expected approximation properties). The idea now is to only add the normal edge bubble."

According to John, Linke, Merdon, Neilan, and Rebholz [693] (2017) (example 6.3), " the velocity space in the Bernardi-Raugel element consists of P_1 functions which are enriched with edge bubble functions". The authors also speak of 'reconstructing the test functions' and state: "the results of the method with reconstruction are generally more accurate. In summary, the use of an appropriately reconstructed test function in the Bernardi– Raugel pair of spaces led to a clear improvement of the accuracy of the computed results compared with the standard method."

In [91] the authors use the so-called Extended Bernardi-Raugel (EBR) element which they call $(P_1 + B_3^F)^3 - P_0$ "which is a cheap and stable pairing with few degrees of freedom (DOFs) per cell".

In John [688] we read: "The pair $\boldsymbol{P}_1^{BR} \times P_0$ is used only occasionally for academic purposes. It is remarked in Bernardi and Raugel [87] that both, the construction of the velocity space and the proof of the inf-sup condition, can be extended to regular families of triangulations consisting of quadrilaterals. This result was used in proving the discrete inf-sup condition for spaces with discontinuous pressure on quadrilateral grids.".

`https://defelement.com/elements/bernardi-raugel.html`

In **zhan23** we read: " The Bernardi–Raugel finite element is a bubble enriched P1 -P0 finite element for the Stokes equations, where three P2 edge-bubbles and four P3 face-bubbles are added to the P1 velocity space on each triangle and each tetrahedron respectively. In the new macro-bubble enriched finite element method, the velocity space is also enriched by three P2 but multi-piece bubbles on each triangle. The divergence of such a P2 bubble function is not a piecewise P1 function but a one-piece P0 function on the triangle. In 3D, four P1 macro-bubbles are added to the P1 velocity space where the divergence of such a macro-bubble function is also a one-piece P0 function. The macro-bubble enriched P1–P0 finite element is shown stable and quasi-optimal in solving the Stokes equations."

10.5.15 The Scott-Vogelius pair

It originates in Larkin Ridgway Scott and Michael Vogelius. "Conforming finite element methods for incompressible and nearly incompressible continua". In: *Lectures in Applied Mathematics* 22.2 (1985). DOI: xxxx (1985).

In Remark 9 (p.29) of Auricchio, Veiga, Brezzi, and Lovadina [34] (2017) we find:

> We also remark that the discontinuous pressure version of the Hood–Taylor element typically results in an unstable method. However, stability can be recovered by imposing certain restrictions on the mesh for $k \geq 3$ (see [1427]; [1223]), or by taking advantage of suitable stabilization procedures for $k \geq 1$ (see Mansfield, 1982; Boffi, 1995).

In Farrell, Mitchell, Scott, and Wechsung [411] (2021) we find:

> The Scott-Vogelius element is given by choosing continuous piecewise polynomials of degree k for the velocity and discontinuous piecewise polynomials of degree $k - 1$ for the pressure. While this clearly implies that $\nabla \cdot V_h \in Q_h$, inf-sup stability of the Scott–Vogelius element is more delicate, and is a topic of ongoing research. In two dimensions, Scott & Vogelius proved [1223] that the element is inf-sup stable for $k \geq 4$ if the mesh does not have nearly singular vertices. In three dimensions, it was proven more recently in [1510] that the element is stable for $k \geq 6$ on uniform meshes. The stability on general tetrahedral meshes continues to be an open question.

> On barycentrically refined meshes, however, the pair is known to be stable for polynomial order $k = d$, see [48, Section 4.6] for the 2D case and [1509] for the 3D case. If one is willing to consider the more complicated Powell–Sabin split, the order can be reduced further to $k = d - 1$ [1512, 1513]. The two refinement patterns are shown for the two dimensional case in Figure 1 [see below]. In this work we will consider the case of $k \geq d$ on barycentrically refined meshes, but the arguments apply mutatis mutandis to the Powell–Sabin split.
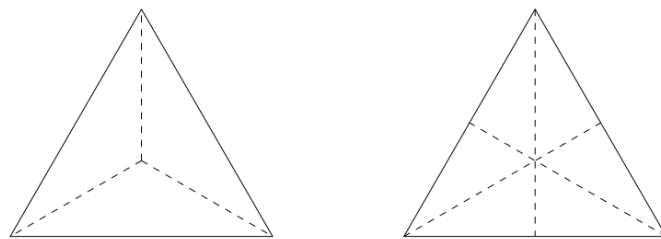
Barycentrically refined triangle (also known as Alfeld split) on the left, and Powell–Sabin split on the right.

Taken from Farrell, Mitchell, Scott, and Wechsung [411] (2021).

Case, Ervin, Linke, and Rebholz [229] (2011) state:

> The SV element pair is not yet very well known, and so we now give a brief description of it. In essence, the SV pair is the same as the Taylor-Hood pair except that the pressure space is discontinuous and either (i) for $k \geq d$, the mesh is a barycenter refinement of a regular mesh, or (ii) for $k = 2, d = 3$, the mesh is formed from a barycenter refined mesh by connecting the barycenter nodes (i.e., a Powell–Sabin tetrahedralization). In short, polynomials of degree k and $k-1$ are used to approximate the velocity and pressure spaces, respectively, and the mesh $\mathcal{T}_h$ that is used must be derived from a regular triangulation (tetrahedralization) of Ω, where each element is refined as stated above. With these mesh constructions, it was proved by Zhang in [42, 44] that the SV elements are LBB stable,

and, consequently, also have optimal approximation properties. It is well known that the TH pair is LBB stable and admits optimal approximation properties for these cases as well [9]. We will restrict our definition of SV elements to these cases where they are LBB stable.

On page 112 of John [688] we read:

The Scott–Vogelius finite element considers still P_k/P_{k-1}^{disc}, $k \geq d$, but on special meshes, which allow to show the satisfaction of the discrete inf-sup condition. [...] This pair of finite element spaces P_k/P_{k-1}^{disc} are weakly divergence-free, which is a desirable property. [...] The fulfillment of the discrete inf-sup condition was proved already in Scott and Vogelius [1223] (1985) in the two-dimensional case for $k \geq 4$ if there is no so-called singular vertex in the mesh. An internal vertex is said to be singular if edges which meet at the vertex fall onto two straight lines.

The basic idea to overcome this problem consists in using meshes without singular vertices. To this end, so-called barycentric-refined grids are constructed. Starting from any admissible triangular mesh, new edges are introduced by connecting all vertices of a mesh cell with the barycenter of this mesh cell. This step creates smaller triangles. On barycentric-refined meshes, the P_k/P_{k-1}^{disc}, $k = 2, 3$ pair of finite element spaces was shown to satisfy the discrete inf-sup condition in Qin's phd thesis (1994). [...]

The use of the P_2/P_1^{disc} pair of finite element spaces on barycentric-refined meshes can be found occasionally in the literature, in particular to demonstrate the advantages of using pairs of finite element spaces which provide weakly divergence-free velocity solutions, e.g. see **john15** and refs therein.
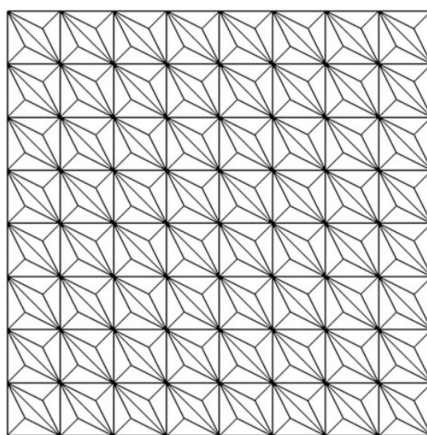
Barycentric-refined simplicial grid on the unit square

I hereafter present my own internal numbering for the mesh above (used in STONE 120 for example). The quadrilateral has been cut once along a diagonal, and then an Alfeld split is used, thereby dividing the square in 6 triangles:

(tikz_sv.tex)

425
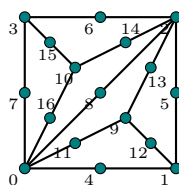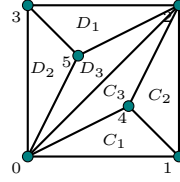
See also John, Linke, Merdon, Neilan, and Rebholz [693] (2017) in which the $P_2 \times P_1$, Scott-Vogelius ($P_2 \times P_{-1}$), Bernardi-Raugel, and $P_2^+ \times P_{-1}$ elements are compared for a thermo-mechanically driven convection problem in a triangle (see STONE 51, although I use the $P_1^+ \times P_1$ element in this stone).

Example 3.73 ($P_k/P_{k-1}^{\mathrm{disc}}$, $k \geq 2$, on a Special Macro Cell) Consider the pair of spaces $P_k/P_{k-1}^{\mathrm{disc}}$, $k \geq 2$, on the grid shown in Fig. 3.5, in particular in the macro cell which is surrounded boldly. The diagonals of the macro cell should be parallel to the coordinate axes. The pair $P_k/P_{k-1}^{\mathrm{disc}}$, $k \geq 2$, is called Scott–Vogelius finite element, see Scott and Vogelius (1985).

Let $\boldsymbol{v}^h \in P_k$, then $\nabla \cdot \boldsymbol{v}^h \in P_{k-1}^{\mathrm{disc}} = Q^h$. Hence, Remark 3.56 gives that discretely divergence-free finite element velocities from $P_k/P_{k-1}^{\mathrm{disc}}$ are even weakly divergence-free, which is a desirable property. However, it turns out that $P_k/P_{k-1}^{\mathrm{disc}}$ does not fulfill the discrete inf-sup condition on the mesh presented in Fig. 3.5.
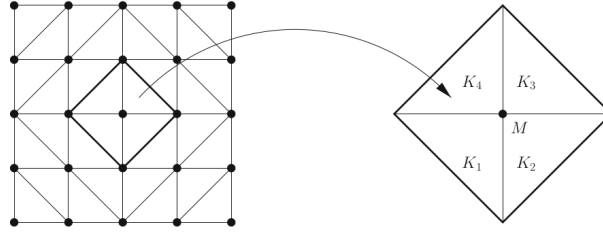
Fig. 3.5 Grid for $P_k/P_{k-1}^{\mathrm{disc}}$, $k \geq 2$, finite element

Taken from John [688, p70].

In Betteridge, Farrell, and Ham [91] (2021) this element is used in its $(P_3)^2 - P_2^{\mathrm{disc}}$ form.

Note that some have proposed to use an incenter-based refinement instead of a barycenter refinement since it lead to less pronounced aspect ratios[16]. In geometry, the incenter of a triangle is a triangle center, a point defined for any triangle in a way that is independent of the triangle's placement or scale. The incenter may be equivalently defined as the point where the internal angle bisectors of the triangle cross or as the point equidistant from the triangle's sides.

Given the coordinates of the three vertices of a triangle ABC, the coordinates of the incenter O are

$$x_O = \frac{ax_A + bx_B + cx_C}{a + b + c} \qquad y_O = \frac{ay_A + by_B + cy_C}{a + b + c}$$

where a, b and c are the side lengths opposite vertex A, B and C.

Note that Zhang [1511] (2024) proposes a solution to the unstable issue:

> We show that the discrete velocity solution converges at the optimal order when solving the steady state Stokes equations by the $\boldsymbol{P}_k \times P_{-(k-1)}$ mixed finite element method for $k \geq 4$ on 2D triangular grids or $k \geq 6$ on tetrahedral grids, even in the case the inf-sup condition fails. By a simple L_2-projection of the discrete P_{k-1} pressure to the space of continuous P_{k-1} polynomials, we show this post-processed pressure solution also converges at the optimal order. Both 2D and 3D numerical tests are presented, verifying the theory.

---

[16]The Scott-Vogelius Method for Stokes Problem on Anisotropic Meshes, K Kean, M Neilan, M Schneier, `https://doi.org/10.48550/arXiv.2109.14780`

Rather interestingly, we find in Terrel, Scott, Knepley, Kirby, and Wells [1330] (2012) a penalty-approach:

> Other solutions to deal with the LBB condition include the Uzawa iteration method and penalty methods. A combination of these two approaches results in the iterated penalty method presented in Scott and Vogelius (1985). Let $r \in \mathbb{R}$ and $\rho \in \mathbb{R}^+$ be prescribed parameters. We wish to find $u^n \in V_h$ such that
>
> $$a(u^n, v) + r(\nabla \cdot u^n, \nabla \cdot v) = (f, v) - (\nabla \cdot v, \nabla \cdot w^n) \qquad \forall\, v \in V_h,$$
>
> where $w^{n+1} = w^n + \rho u^n$. The pressure may be recovered from the auxiliary field w via $p = \nabla \cdot w - C$, where C is an arbitrary constant (since the pressure field is only determined up to an arbitrary constant). When computing the error in p, we subtract the average of $\nabla \cdot w$ to account for C. The algorithm initially assumes $w^0 = 0$, and then solves [the equation above] and updates w. The process is repeated until $\|u^{n+1} - u^n\| < \epsilon$, where ϵ is a prescribed tolerance. This method involves only one function space, but it requires a higher-order continuous element ($q \geq 4$) and it solves the divergence-free criterion exactly. The iteration count and accuracy are dependent upon the penalty parameters ρ and r. The implementation of this formulation is presented in [the following figure]:

Python code

```python
# Define function space
V = VectorFunctionSpace(mesh, "Lagrange", U_order)

# Define trial and test functions
u = TrialFunction(V)
v = TestFunction(V)

# Define auxiliary function and parameters
w = Function(V);
rho = 1.0e3
r = -rho

# Define the variational problem
a = inner(grad(u), grad(v))*dx + r*div(u)*div(v)*dx
L = inner(f, v)*dx + inner(div(w), div(v))*dx
pde = VariationalProblem(a, L, bc0)

# Iterate to fix point
iters = 0; max_iters = 100; U_m_u = 1
while iters < max_iters and U_m_u > 1e-8:
    U = pde.solve()
    w.vector().axpy(rho, U.vector())
    if iters != 0:
        U_m_u = (U.vector() - u_old_vec).norm("l2")
    u_old_vec = U.vector().copy()
    iters += 1
```

https://defelement.com/elements/scott-vogelius.html

10.5.16   The BDM (Brezzi-Douglas-Marini) pair

This element is mentioned in Kanschat book [708], section 4.2.14. It also exists for quads see section 4.2.39 in the same book. It is mentioned in Chen [241] (1993), also check the book by Brezzi and Fortin [157]. It is well described in Kanschat [710]. There is an entire chapter (14) of Ern and Guermond [404] dedicated to H(div) and section 14.5.1 to BDM elements. Check section 4.1.1 of [34] for triangles and quads.

https://defelement.com/elements/brezzi-douglas-marini.html

- In Logg, Mardal, and Wells [852] we read:

  The Brezzi-Douglas-Marini element was introduced by Brezzi, Douglas and Marini in two dimensions (for triangles) in Brezzi, Douglas, and Marini [161] (1985). The element can be viewed as an alternative to the Raviart-Thomas element using a complete polynomial space. It

was later extended to three dimensions (tetrahedra, prisms and cubes) in Nédélec [987] (1986) and Brezzi, Douglas, Durán, and Fortin [160] (1987). The definition given here is based on that of Nédélec [987] (1986).

The Brezzi-Douglas-Marini element was introduced for mixed formulations of second-order elliptic equations. However, it is also useful for weakly symmetric discretizations of the elastic stress tensor; see Farhloul and Fortin (1997); Arnold et al. (2007).
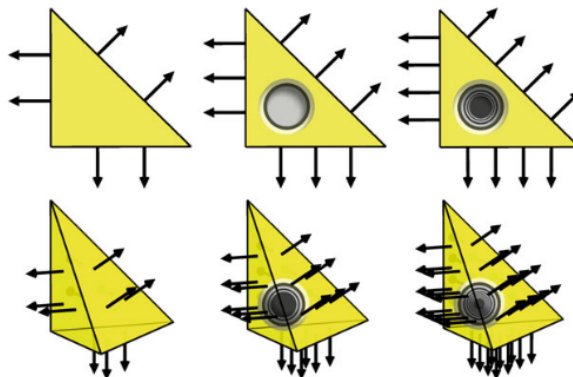
Taken from [852].

The dimension of BDM_q is $(q+1)(q+2)$ for a triangle and $\frac{1}{2}(q+1)(q+2)(q+3)$ for a tetrahedron. Check book for definition.

A slight modification of the Brezzi-Douglas-Marini element constrains the element space $\mathcal{V}$ by only allowing normal components on the boundary of polynomial degree $q-1$ (rather than the full polynomial degree q). Such an element was suggested on rectangles by Brezzi, Douglas, Fortin, and Marini [163] (1987), and the triangular analogue was given in Brezzi and Fortin [157]. In similar spirit, elements with differing degrees on the boundary suitable for varying the polynomial degree between triangles were derived in Brezzi, Douglas, and Marini [162] (1985).

- On the defelement website[17] we find a lot of information. Note that the mapping is set to 'contravariant Piola'.

  I still need to understand and write about this!

It belongs to the categories 'Vector-valued elements', and 'H(div) conforming elements'
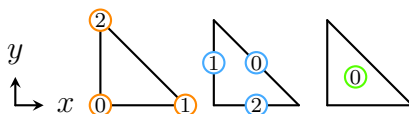
Taken from DefElement `https://defelement.org/img/ref-triangle.html`. I have altered the font size. Orange: nodes; Blue: edges.

I reproduce below the figures and basis functions pertaining to the Degree 1 triangle, but the site also shows Degree 2 triangle, Degree 1 & 2 tetrahedron, and so-called Lagrange variants.
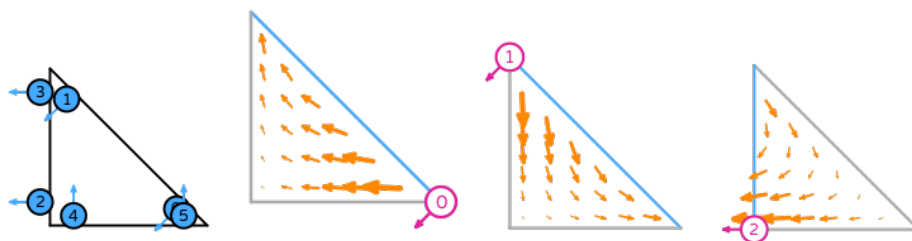
---

[17]`https://defelement.org/elements/brezzi-douglas-marini.html`
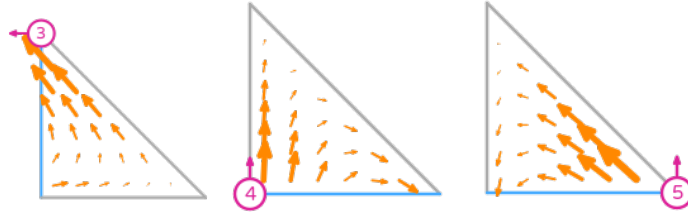
Pink: degrees of freedom.

$\mathcal{V}$ is spanned by

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} x \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x \end{pmatrix}, \begin{pmatrix} y \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ y \end{pmatrix}$$

with

- DOF #0 is associated with edge 0 of the reference element with $\vec{\mathcal{N}_0}$ basis function.
- DOF #1 is associated with edge 0 of the reference element with $\vec{\mathcal{N}_1}$ basis function.
- DOF #2 is associated with edge 1 of the reference element with $\vec{\mathcal{N}_2}$ basis function.
- DOF #3 is associated with edge 1 of the reference element with $\vec{\mathcal{N}_3}$ basis function.
- DOF #4 is associated with edge 2 of the reference element with $\vec{\mathcal{N}_4}$ basis function.
- DOF #5 is associated with edge 2 of the reference element with $\vec{\mathcal{N}_5}$ basis function.

and

$$
\begin{aligned}
\vec{\mathcal{N}_0} &= \begin{pmatrix} -4x \\ 2y \end{pmatrix} \\
\vec{\mathcal{N}_1} &= \begin{pmatrix} 2x \\ -4y \end{pmatrix} \\
\vec{\mathcal{N}_2} &= \begin{pmatrix} 4x + 6y - 4 \\ -2y \end{pmatrix} \\
\vec{\mathcal{N}_3} &= \begin{pmatrix} -2x - 6y + 2 \\ 4y \end{pmatrix} \\
\vec{\mathcal{N}_4} &= \begin{pmatrix} 2x \\ -6x - 4y + 4 \end{pmatrix} \\
\vec{\mathcal{N}_5} &= \begin{pmatrix} -4x \\ 6x + 2y - 2 \end{pmatrix}
\end{aligned}
$$

- In [163] we read:

  The object of this paper is to present families of rectangular mixed finite éléments that are derived from the elements of [161] in two space variables and of [160] in three space variables.

10.5.17 The Divergence-free nonconforming $P_1^{NC} \times P_0$ pair

It belongs to the Crouzeix-Raviart family. The midside nodes are used as degrees of freedom for the velocities. It is mentioned in Section 6.3 of Boffi, Brezzi, and Fortin [114] (2008):

[...] It is exactly divergence free. Another important feature of this element is that it can be seen as a "mass conservation" scheme. The present element has been generalized to second order in Fortin and Soulie [430] (1983). It must also be said that coerciveness may be a problem for the $P_1^{NC} \times P_0$ element, as it does not satisfy the discrete version of Korn's inequality. This issue has been deeply investigated and clearly illustrated in Arnold [28] (1993).
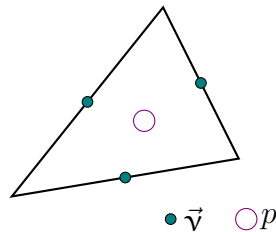
At page 170 of [136] it is stated that "an analogous quadrilateral element was developed and studied by Rannacher and Turek [1112] (1992)" (see Section 10.4.9).

In Boffi, Brezzi, and Fortin [115] we find:

> We consider the classical (almost[18]) stable nonconforming triangular element introduced in Crouzeix and Raviart [309], in which mid-side nodes are used as degrees of freedom for the velocities. This generates a piecewise linear nonconforming approximation; pressures are taken constant on each element. It is also possible to build a three-dimensional version of this element, using mid-face nodes as degrees of freedom.
>
> ..
>
> It must also be recalled that coercivity is a problem for the $P_1^{NC} \times P_0$ element. The trouble is that the bilinear form (8.2.1) is not coercive on the nonconforming space V_h and we do not have the discrete version of Korn's inequality.

It is also mentioned in John [688], appendix B.3, example B.43, in 2D and 3D, in Brezzi and Fortin [157] (example 8.1), and studied extensively in John [687] (1998).

---

[18]What does that mean?!
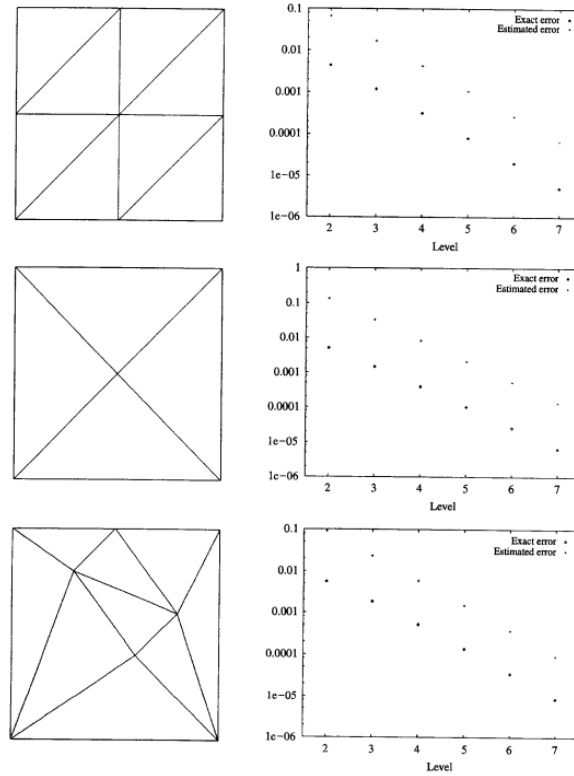
In John, Linke, Merdon, Neilan, and Rebholz [693] (2017) the authors show results obtained with this element (fig 6) but also explain that these are obtained with so-called reconstructed test functions.

10.5.18 Mardal-Tai-Winther element pair

pair_mtw.tex

- In Logg, Mardal, and Wells [852] we read:

  The Mardal-Tai-Winther element was introduced in Mardal, Tai, and Winther [883] (2002) as a finite element suitable for the velocity space for both Darcy and Stokes flow in two dimensions. In the Darcy flow equations, the velocity space only requires $H(\mathrm{div})$-regularity. Moreover, discretizations based on H^1-conforming finite elements are typically not stable. On the other hand, for the Stokes equations, the velocity space does stipulate H^1-regularity. The Mardal-Tai-Winther element is $H(\mathrm{div})$-conforming, but H^1-nonconforming. The element was extended to three dimensions in Tai and Winther [1315] (2006).
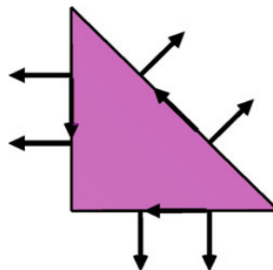
Illustration of the Mardal-Tai-Winther element. The degrees of freedom are two moments of the normal component on each facet and

one moment of the tangential component on each facet. In this figure, the moments of normal components are illustrated by point

evaluation of normal components.

- On the DefElement website[19] we find that it belongs to the 'Vector-valued elements' and 'H(div) conforming elements' categories. It is also associated with the 'contravariant Piola' mapping. ⟨learn,exp⟩ It is available on triangles and tetrahedra. On the site we find a degree 3 triangle and a degree 3 tetrahedron.
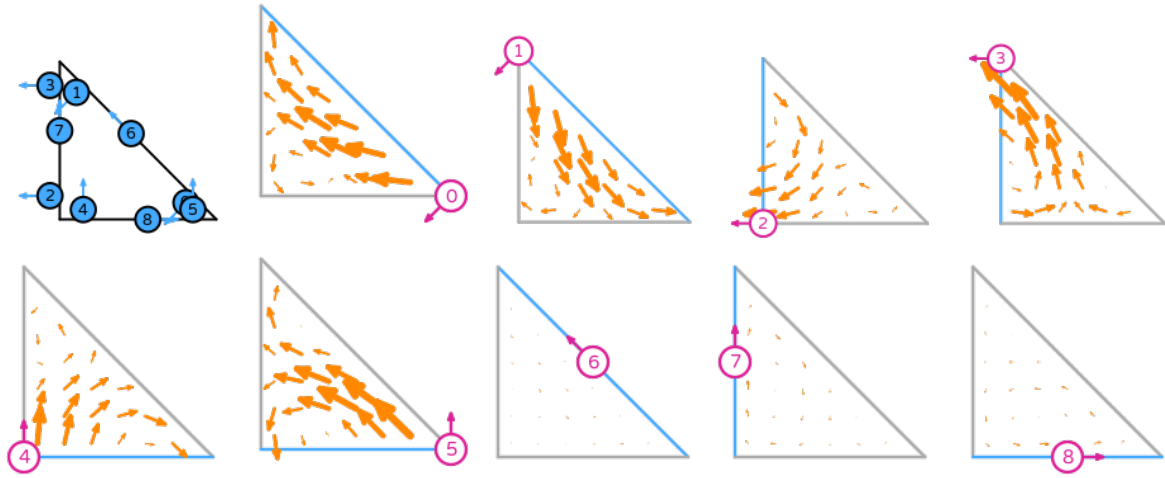
Pink: degrees of freedom.

$\mathcal{V}$ is spanned by

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ 0 \end{pmatrix}, \begin{pmatrix} y \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ x \end{pmatrix}, \begin{pmatrix} 0 \\ y \end{pmatrix}, \begin{pmatrix} x(x+2y) \\ y(-2x-y) \end{pmatrix}, \begin{pmatrix} x(-x^2+2x+3y^2) \\ y(3x^2-4x-y^2) \end{pmatrix}, \begin{pmatrix} x(2xy+x+3y \\ y(-2x-y) \end{pmatrix}$$

with

- DOF #0 is associated with edge 0 of the reference element with $\vec{\mathcal{N}}_0$ basis function.
- DOF #1 is associated with edge 0 of the reference element with $\vec{\mathcal{N}}_1$ basis function.
- DOF #2 is associated with edge 1 of the reference element with $\vec{\mathcal{N}}_2$ basis function.
- DOF #3 is associated with edge 1 of the reference element with $\vec{\mathcal{N}}_3$ basis function.
- DOF #4 is associated with edge 2 of the reference element with $\vec{\mathcal{N}}_4$ basis function.
- DOF #5 is associated with edge 2 of the reference element with $\vec{\mathcal{N}}_5$ basis function.
- DOF #6 is associated with edge 0 of the reference element with $\vec{\mathcal{N}}_6$ basis function.
- DOF #7 is associated with edge 1 of the reference element with $\vec{\mathcal{N}}_7$ basis function.
- DOF #8 is associated with edge 2 of the reference element with $\vec{\mathcal{N}}_8$ basis function.

---

[19]https://defelement.org/elements/mardal-tai-winther.html

and

$$\vec{\mathcal{N}}_0 = \begin{pmatrix} x(30x^2 + 144xy - 57x + 126y^2 - 138y + 23) \\ y(-90x^2 - 144xy + 114x - 42y^2 + 69y - 25) \end{pmatrix}$$

$$\vec{\mathcal{N}}_1 = \begin{pmatrix} x(-42x^2 - 144xy + 69x - 90y^2 + 114y - 25) \\ y(126x^2 + 144xy - 138x + 30y^2 - 57y + 23) \end{pmatrix}$$

$$\vec{\mathcal{N}}_2 = \begin{pmatrix} -24x^3 - 24x^2y + 24x^2 + 36xy^2 - 24xy + 4x + 6y - 4 \\ 2y(36x^2 + 12xy - 24x - 6y^2 + 6y - 1) \end{pmatrix}$$

$$\vec{\mathcal{N}}_3 = \begin{pmatrix} 48x^3 + 120x^2y - 66x^2 + 36xy^2 - 60xy + 16x - 6y + 2 \\ 2y(-72x^2 - 60xy + 66x - 6y^2 + 15y - 7) \end{pmatrix}$$

$$\vec{\mathcal{N}}_4 = \begin{pmatrix} 2x(6x^2 - 12xy - 6x - 36y^2 + 24y + 1) \\ -36x^2y + 24xy^2 + 24xy - 6x + 24y^3 - 24y^2 - 4y + 4 \end{pmatrix}$$

$$\vec{\mathcal{N}}_5 = \begin{pmatrix} 2x(6x^2 + 60xy - 15x + 72y^2 - 66y + 7) \\ -36x^2y - 120xy^2 + 60xy + 6x - 48y^3 + 66y^2 - 16y - 2 \end{pmatrix}$$

$$\vec{\mathcal{N}}_6 = \begin{pmatrix} 3x(-2x^2 - 8xy + 3x - 6y^2 + 6y - 1) \\ 3y(6x^2 + 8xy - 6x + 2y^2 - 3y + 1) \end{pmatrix}$$

$$\vec{\mathcal{N}}_7 = \begin{pmatrix} 6x(-2x^2 - 4xy + 3x + 2y - 1) \\ 6y(6x^2 + 4xy - 6x - y + 1) \end{pmatrix}$$

$$\vec{\mathcal{N}}_8 = \begin{pmatrix} 6x(4xy - x + 6y^2 - 6y + 1) \\ 6y(-4xy + 2x - 2y^2 + 3y - 1) \end{pmatrix}$$

- In a presentation online I found the following slide from F. Aznaran et al, given at the FEniCS 2021 conference:

The Mardal–Tai–Winther element

$MTW(K) = \left\{ \Phi \in \mathcal{P}_3(K; \mathbb{R}^2) \,|\, \mathrm{div}\,\Phi \in \mathcal{P}_0(K), (\Phi \cdot \mathbf{n})\,|_{\mathbf{e}} \in \mathcal{P}_1(\mathbf{e}) \,\forall\, \text{edges}\, \mathbf{e} \right\}$

Novelties:
- Discretises $H(\mathrm{div})$ and (nonconforming) $\mathbf{H}^1(\Omega)$
- Stable for both Stokes and Darcy flow
- Locking-free element for (poro)elasticity; discrete Korn inequality
- Divergence-free for Stokes when paired with $DG(0)$
- Discretises the 2D Stokes complex:

$$\mathbb{R} \xrightarrow{\subset} H^2(\Omega) \xrightarrow{\mathrm{curl}} \mathbf{H}^1(\Omega) \xrightarrow{\mathrm{div}} L^2(\Omega) \longrightarrow 0.$$
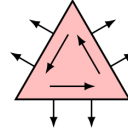
10.5.19 Other FE element pairs

- check Dhatt and Hubert [352] (1986) many flavours of triangles and quads.

- Bercovier-Pironneau element pair, or $P_1 iso P_2$. See Boffi, Cavallini, Gardini, and Gastaldi [116] (2012).

- Johnson–Mercier element [34]

- Arnold–Winther element [34], chapter 3.4.4 of [852]

- Nedelec second-kind `https://defelement.org/elements/nedelec2.html` [987]

10.6 Which velocity-pressure pair for Stokes? (Raviart-Thomas spaces)

10.6.1 What the literature says

This element is substantially more complicated than most of the others in FieldStone so I have proceeded systematically by scanning the literature (i.e. textbooks mostly) for useful definitions and statements. What follows is mostly a collection of relevant/interesting statements.

- There is an entire chapter (14) of Ern and Guermond [404] dedicated to Raviart-Thomas elements and $H(\text{div})$.

- Likewise Chapter 3 of Gatica [469] is dedicated to Raviart-Thomas spaces.

- Conerning the Stokes equations, I have found a reference[20] that uses the stable element RT0-P0 (the lowest order Raviart-Thomas element for velocity and piecewise constant for pressure) of Stokes equations in 2-D. In Lehmann, Lukacova-Medvidova, Kaus, and Popov [807] (2015) we read:

  > As shown in [709] for the isoviscous case, the $RT_0 - P_0$ element in the corresponding numerical scheme resembles the finite difference (FD) staggered grid stencil. Raviart-Thomas elements of order k can, therefore, be considered a higher-order generalization of the FD method working on irregular meshes.

  Note that the authors compare $Q_1 \times P_0$, $Q_2 \times P_{-1}$, $RT_0 - P_0$ and $BDM_1 - P_0$ for the Stokes equation but the two last ones rely on DG.

  Interestingly enough, the RT elements are described in [34] but only used for thermal problems, not Stokes, while other elements are proposed for Stokes. Could it be bc for RT0-P0 on quads we have $\dot{\varepsilon}_{xy}^h = 0$ due to the formulation of the basis functions ? Is this why [807] use it in DG formulation and not CG?

- In Hanert, Legat, and Deleersnijder [561] (2003) we find:

  > $P_1^\perp \times P_0$ symbol denotes an element with normal velocity nodes in the middle of each edge of the triangulation [...]. This element, also called low order Raviart–Thomas element (Raviart and Thomas, 1977 [1115]), is based on flux conservation on elements edges and the resulting scheme is very close to a finite volume scheme.

- It is mentioned in John [688], appendix B.3, example B.45:

  > the normal component of v on each face is a constant. The normal component of functions from RT0 is continuous across faces of the mesh cells.

- In Chen [241] (1993) we read:

  > Arnold and Brezzi [2] showed, for example, that the Raviart-Thomas mixed method of lowest order is equivalent to the usual P_1-nonconforming method modified by augmenting the classical P_1-nonconforming space with P_3-bubbles and then proved that the equivalence is useful not only for implementing the mixed method but also for deriving error estimates.

---

[20]https://lyc102.github.io/ifem/fem/StokesRT0femrate/

- There is an entry on Wikipedia[21] but it is short and not so useful, except for

  In applied mathematics, Raviart–Thomas basis functions are vector basis functions used in finite element and boundary element methods. They are regularly used as basis functions when working in electromagnetics.

- There is also a useful entry on Stackexchange[22]:

  There are quadrilateral Raviart-Thomas elements. In two dimensions, the polynomial space for RT_k is given by $P_{k+1,k} \times P_{k,k+1}$, where

  $$P_{k,l} = \left\{ \sum_{i=0}^{k} \sum_{j=0}^{l} a_{ij} x^i y^j : a_{ij} \in \mathbb{R} \right\}$$

  A typical polynomial for $k = 0$ would thus be

  $$\begin{pmatrix} a_1 + b_1 x \\ a_2 + b_2 y \end{pmatrix}$$

  and for $k = 1$:
  $$\begin{pmatrix} a_1 + b_1 x + c_1 x^2 + d_1 y + e_1 xy + f_2 x^2 y \\ a_2 + b_2 y + c_2 y^2 + d_2 x + e_2 xy + f_2 xy^2 \end{pmatrix}$$

  Hence, $dim(RT_1) = 12$, and in general, $dim(RT_k) = 2(k+1)(k+2)$. This means you need two additional degrees of freedom, which should be located on the interior of the element. (In general, for RT_k you take $k+1$ normal derivatives on each facet, and the remaining degrees of freedom from the interior.)

  For Raviart-Thomas elements, you usually take moments rather than point evaluations, i.e., the remaining conditions come from the conditions

  $$\int_{-1}^{+1} \int_{-1}^{+1} \phi_i(x,y) q_j(x,y) dx dy = \delta_{ij}$$

  where the q_j are a basis of $P_{k-1,k} \times P_{k,k-1}$ (e.g. $1, x, y$ for $k = 1$). To make it easier to get a full nodal basis, the facet degrees of freedom are usually not taken as point evaluation, but also as moment conditions:

  $$\int_{e_m} \phi_i(s)^T \nu_{e_m} q_{m,j}(s) ds$$

  where e_m is one of the four edges, ν_{e_m} is the corresponding outer normal, and for each m, the $q_{m,j}$ form a basis of $P_k(e_m)$ (e.g. 1,x or 1,y for $k = 1$ depending on the edge orientation). Together, these degrees of freedom are unisolvent (i.e., the corresponding system of basis functions is always invertible).

- In Kanschat [710] we find:

  The Raviart-Thomas element of degree $k \geq 0$ on simplices consists of the polynomial space
  $$RT_k = \mathbb{P}_k^d + x\mathbb{P}_k.$$

---

[21]https://en.wikipedia.org/wiki/Raviart-Thomas_basis_functions
[22]https://scicomp.stackexchange.com/questions/20245/raviart-thomas-elements-on-reference-square

For the simplicial Raviart-Thomas element in $\mathbb{R}^d$ there holds

$$dim(RT_k) = (d+1)dim(\mathbb{P}_k) - dim(\mathbb{P}_{k-1})$$

In particular:

$$
\begin{aligned}
dim(RT_k)|_{2D} &= (k+1)(k+3) \\
dim(RT_k)|_{3D} &= \frac{1}{2}(k+1)(k+2)(k+4)
\end{aligned}
$$
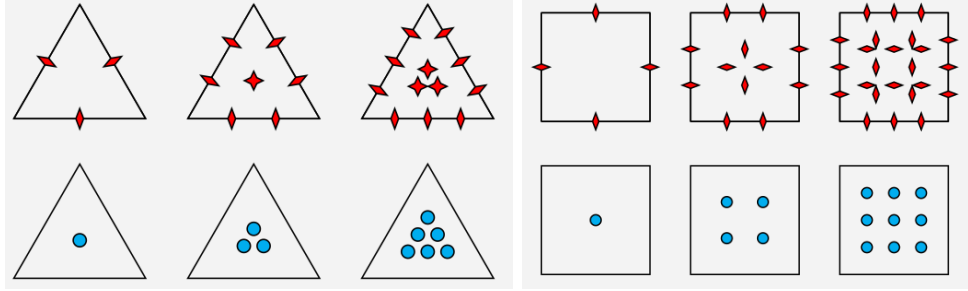
Examples 4.2.10 and 4.2.37 of Kanschat [**710**]. The first members of the Raviart-Thomas family on triangles and quadrilaterals.

The construction principle of the Raviart-Thomas element on simplices as well as on rectangles and cubes can be seen as adding vector polynomials to the velocity space until its divergence is equal to $\mathbb{P}_k$ or $\mathbb{Q}_k$.

The Raviart-Thomas element of degree $k \geq 0$ on the reference cell $\hat{T} = [-1,1]^d$ consists of the polynomial space

$$RT_k(\hat{T}) = \mathbb{Q}_k^d(\hat{T}) + x\mathbb{Q}_k(\hat{T})$$

and we have

$$dim(RT_k) = d(k+1)^{d-1}(k+2)$$

- This element is implemented in deal.II and we read[23]:

  The Raviart-Thomas space is designed to solve problems in which the solution only lives in the space $H^{\mathrm{div}} = \{\boldsymbol{u} \in L_2 : \mathrm{div}(\boldsymbol{u}) \in L_2\}$, rather than in the more commonly used space $H^1 = \{\boldsymbol{u} \in L_2 : \nabla\boldsymbol{u} \in L_2\}$. In other words, the solution must be a vector field whose divergence is square integrable, but for which the gradient may not be square integrable. The typical application for this space (and these elements) is to the mixed formulation of the Laplace equation and related situations, see for example step-20. The defining characteristic of functions in H^{div} is that they are in general discontinuous - but that if you draw a line in 2d (or a surface in 3d), then the normal component of the vector field must be continuous across the line (or surface) even though the tangential component may not be. As a consequence, the Raviart-Thomas element is constructed in such a way that (i) it is vector-valued, (ii) the shape functions are discontinuous, but (iii) the normal component of the vector field represented by each shape function is continuous across the faces of cells.

  Other properties of the Raviart-Thomas element are that (i) it is not a primitive element; (ii) the shape functions are defined so that certain integrals over the faces are either zero or one, rather than the common case of certain point values being either zero or one.

---

[23]https://www.dealii.org/current/doxygen/deal.II/classFE__RaviartThomas.html

We follow the commonly used - though confusing - definition of the "degree" of RT elements. Specifically, the "degree" of the element denotes the polynomial degree of the largest complete polynomial subspace contained in the finite element space, even if the space may contain shape functions of higher polynomial degree. The lowest order element is consequently FE_RaviartThomas(0), i.e., the Raviart-Thomas element "of degree zero", even though the functions of this space are in general polynomials of degree one in each variable. This choice of "degree" implies that the approximation order of the function itself is degree+1, as with usual polynomial spaces.

- Elsewhere in the deal.II library[24] we read[25]

  The Raviart-Thomas polynomials are constructed such that the divergence is in the tensor product polynomial space Q_k. Therefore, the polynomial order of each component must be one order higher in the corresponding direction, yielding the polynomial spaces $(Q_{k+1,k}, Q_{k,k+1})$ and $(Q_{k+1,k,k}, Q_{k,k+1,k}, Q_{k,k,k+1})$ in 2d and 3d, respectively.

  The Raviart-Thomas element is used in Step-20 (`https://www.dealii.org/current/doxygen/deal.II/step_20.html`) and Step-21 (`https://www.dealii.org/current/doxygen/deal.II/step_21.html`).

- In John's book [688] we find:

  Raviart–Thomas finite elements are a class of vector-valued finite elements that approximate the space $H(\mathrm{div}, \Omega) = \{\boldsymbol{v} \in L^2(\Omega) : \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\}$. Details of their definition and important properties can be found, e.g., in page 84 of Boffi, Brezzi, and Fortin [115] (2013).

  Consider a simplicial triangulation with mesh cells $\{K\}$ and let $\boldsymbol{P}_k(K) = (P_k(K))^d$, $k \geq 0$. Then, the following local polynomial space is defined directly on K:

  $$RT_k(K) = \{\boldsymbol{v} \in (\boldsymbol{P}_k(K) + \boldsymbol{x}P_k(K)) : \boldsymbol{v} \cdot \boldsymbol{n}_{\partial K} \in R_k(\partial K)\}, \qquad k \geq 0,$$

  where

  $$R_k(\partial K) = \{\phi \in L^2(\partial K) : \phi|_E \in P_k(E) \text{ for all faces } E \subset \partial K\}$$

  It is noted in Boffi, Brezzi, and Fortin [115] (remark 2.3.1) that this definition of $RT_k(K)$ is different than the original definition in Raviart and Thomas [1115] (1977). In particular, the Raviart-Thomas element of lowest order is given by

  $$RT_0(K) = \{\boldsymbol{v} \in (\boldsymbol{P}_0(K) + \boldsymbol{x}P_0(K)) : \boldsymbol{v} \cdot \boldsymbol{n}_{\partial K} \in R_0(\partial K)\}$$

  i.e., it is linear on K. Its local functionals are $\boldsymbol{v} \cdot \boldsymbol{n}_{E_i}$, $i = 1, ..d + 1$. A function from $RT_0(K)$ can be written in the form

  $$\boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{a} + b\boldsymbol{x}, \quad \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}, \boldsymbol{x} \in K. \qquad B.21$$

  A face $E \subset \partial K$ is a hyperplane that can be represented in the form

  $$\boldsymbol{x} \cdot \boldsymbol{n}_E = c, \qquad c \in \mathbb{R}, \forall \boldsymbol{x} \in E.$$

  Inserting this representation in (B.21) yields

  $$\boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{n}_E = \boldsymbol{a} \cdot \boldsymbol{n}_E + b\boldsymbol{x} \cdot \boldsymbol{n}_E = \boldsymbol{a} \cdot \boldsymbol{n}_E + bc = const \qquad \forall \boldsymbol{x} \in E.$$

  Thus, the normal component of $\boldsymbol{v}$ on each face is a constant.

---

[24]`https://www.dealii.org/current/doxygen/deal.II/classPolynomialsRaviartThomas.html`
[25]Note that at the time deal.II only supported quads.

- On the defelement site[26] we find that it is defined on triangles, tetrahedra, quadilaterals and hexahedra. with the number of dofs defined as follows:

  - triangle: $k(k+2)$
  - tetrahedron: $k(k+1)(k+3)/2$
  - quadrilateral: $2k(k+1)$
  - hexahedron: $3k^2(k+1)$

- In Brezzi and Fortin [157] (1991) we read:

  On triangles we know that RT_0 is a space of dimension 3 containing polynomials of the form

  $$\begin{aligned} q_1(x,y) &= a + cx \\ q_2(x,y) &= b + cy \end{aligned}$$

  a) RT_0 c) RT_1
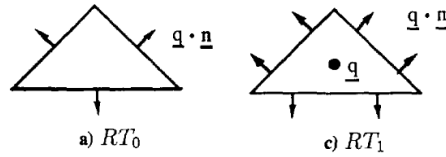
  The simplest cases of three-dimensional elements are depicted here:

  a) RT_0 c) RT_1
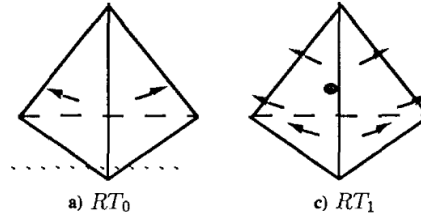
  We now consider the extension of the previous construction to rectangular elements In the present case, the use of a reference element is essential and we shall build our space on $\hat{K} =]-1, 1[^n$. Contrarily to the simplicial case, it will be simpler here to first introduce the approximations of Raviart and Thomas. Let us thus define, as in the previous section,

  $$RT_{[k]} = (Q_k)^n + \underline{x} Q_k$$

  It can be checked that one has

  $$\begin{aligned} RT_{[k]}|_{n=2} &= P_{k+1,k} \times P_{k,k+1} \\ RT_{[k]}|_{n=3} &= P_{k+1,k,k} \times P_{k,k+1,k} \times P_{k,k,k+1} \end{aligned}$$

  and

  $$dim(RT_{[k]}|_{n=2}) = 2(k+1)(k+2)$$
  $$dim(RT_{[k]}|_{n=3}) = 3(k+1)^2(k+2)$$

- In Larson and Bengzon [796] we read:

---

[26]https://defelement.com/elements/raviart-thomas.html

Not all finite elements are scalar and there are also vector valued elements. As the name suggests, vector valued elements are used to approximate vector valued functions. One such element is the Raviart-Thomas element, which is designed to mimic the Hilbert space

$$H(\mathrm{div}; \Omega) = \{v \in [L^2(\Omega)]^2 : \nabla \cdot v \in L^2(\Omega)\}$$

That is, the space of all vectors with bounded divergence. A simple application of Green's formula shows that all such functions must have continuous normal components, which is the basic design feature of the Raviart-Thomas element.

Actually there is a whole family of Raviart Thomas elements, but we shall only study the simplest of them called the RT0 element. On a triangle K the polynomial space for RT_0 is given by $P = [P_0(K)]^2 + [x_1, x_2]^T P_0(K)$, that is, all vectors of the form

$$v = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + b \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

for some coefficients a_1 , a_2 , and b.

If the normal is chosen consistently on each edge in the mesh, then by construction is the RT_0 shape functions are continuous in the normal direction across the edge of any two adjacent elements. This ensures that the RT_0 functions belong to $H(\mathrm{div}, \Omega)$. The isoparametric map can not be used for RT_0 elements, since the divergence is not preserved by this mapping.

The Raviart-Thomas elements typically find application in acoustics and elasticity.
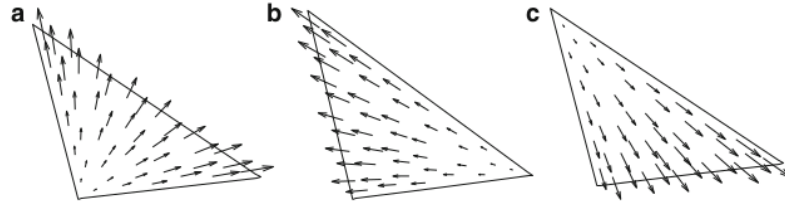
Fig. 8.9 RT_0 shape functions on a *triangle*. (a) $S_1^{RT_0}$, (b) $S_2^{RT_0}$, and (c) $S_3^{RT_0}$

- In Logg, Mardal, and Wells [852] we find:

The Sobolev space $H(\mathrm{div})$ consists of vector fields for which the components and the weak divergence are square-integrable. This is a weaker requirement than for a d-vector field to be in $[H^1]^d$ (for $d \geq 2$). This space naturally occurs in connection with mixed formulations of second-order elliptic problems, porous media flow, and elasticity equations. For a finite element family to be $H(\mathrm{div})$-conforming, each component need not be continuous, but the normal component must be continuous. In order to ensure such continuity, the degrees of freedom of $H(\mathrm{div})$-conforming elements usually include normal components on element facets. The two main families of $H(\mathrm{div})$-conforming elements are the Raviart-Thomas and Brezzi-Douglas-Marini elements.

[...]

The Raviart–Thomas element was introduced by Raviart and Thomas (1977) [1115]. It was the first element to discretize the mixed form of second-order elliptic equations on triangles. Its element space $\mathcal{V}$ is designed so that it is the smallest polynomial space $\mathcal{V} \subset \mathcal{P}_q(T)$, for $q = 1, 2, ...$, from which the divergence maps onto $\mathcal{P}_{q-1}(T)$.

Shortly thereafter, it was extended to tetrahedra and boxes by Nédélec [988] (1980). It is therefore sometimes referred to as the Raviart-Thomas-Nédélec element.
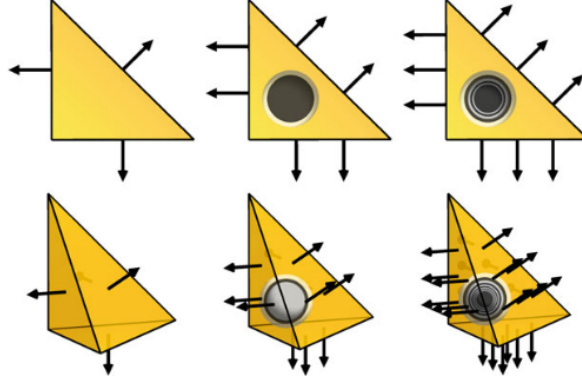
[...]

As an example, the lowest degree Raviart-Thomas space on triangles is a three-dimensional space and consists of vector fields of the form

$$v(x) = \alpha + \beta x$$

where α is a vector-valued constant, and β is a scalar constant.

The dimension of RT_q is $q(q+2)$ for a triangle and $\frac{1}{2}q(q+1)(q+3)$ for a tetrahedron.

- In Auricchio, Veiga, Brezzi, and Lovadina [34], we find several examples of efficient mixed finite element methods, focusing our attention mostly on the thermal problem and on the Stokes equation. We read :

  Throughout this section, we will always suppose that the domain $\Omega \subset \mathbb{R}^2$, on which the thermal problem is posed, is decomposed by means of a triangular mesh $\mathcal{T}_h$ with mesh size h. Moreover, we define $\mathcal{E}_h$ as the set of all the edges of the triangles in $\mathcal{T}_h$. We now introduce the simplest triangular element proposed for thermal problems. For the discretization of the thermal flux $\boldsymbol{q}$, we take the so-called lowest order Raviart–Thomas element (RT_0 element), presented in Raviart and Thomas, 1977 [1115], accordingly, the approximated flux $\boldsymbol{q}^h$ is described as a piecewise linear (vectorial) field such that

  (i) the normal component $\boldsymbol{q}^h \cdot \boldsymbol{n}$ is constant on each edge e of $\mathcal{E}_h$;

  (ii) the normal component $\boldsymbol{q}^h \cdot \boldsymbol{n}$ is continuous across each edge e of $\mathcal{E}_h$;

  To approximate the temperature, we simply use piecewise constant functions in each element (P_0 element). On the generic triangle $T \in \mathcal{T}_h$, a set of element degrees of freedom for $\boldsymbol{q}^h$ is given by its three normal fluxes on the edges of the triangle, that is,

  $$\int_e \boldsymbol{q}^h \cdot \boldsymbol{n} \, ds \qquad \forall e \text{ edge of } T$$

  Therefore, the space for the element approximation of $\boldsymbol{q}$ has dimension 3 and a basis is obtained by considering the (vectorial) shape functions

  $$\boldsymbol{N}_k^{\boldsymbol{q}}(x,y) = \frac{1}{2\,\text{Area}(T)} \left\{ \begin{array}{c} x - x_k \\ y - y_k \end{array} \right\} \qquad k = 1, 2, 3$$

Here, $\{x_k, y_k\}^T$ denotes the position vector of the k-th vertex (local numbering) of the triangle T.

We also remark that, because of the equation above, $\boldsymbol{q}^h$ can be locally described by

$$\boldsymbol{q}^h = \boldsymbol{p}_0 + p_0 \left\{ \begin{array}{c} x \\ y \end{array} \right\} = \left\{ \begin{array}{c} a_0 + p_0 x \\ b_0 + p_0 y \end{array} \right\}$$

where $a_0, b_0, p_0 \in \mathbb{R}$.

As far as the approximated temperature (θ) is concerned, an element basis for θ^h is given by the shape function

$$N^\theta(x, y) = 1$$

We now present the extension to higher orders of the $RT_0 - P_0$ method just described (cf. Raviart and Thomas, 1977). Given an integer $k \geq 1$ and using the definition introduced in Nedelec, 1980 [988]; for the flux $\boldsymbol{q}^h$, we take a field such that (RT_k element) on each triangle T of $\mathcal{T}_h$, we have

$$\boldsymbol{q}^h = \boldsymbol{p}_k(x, y) + p_k(x, y) \left\{ \begin{array}{c} x \\ y \end{array} \right\}$$

where $\boldsymbol{p}_k(x, y)$ is a vectorial polynomial of degree at most k and $p_k(x, y)$ is a scalar polynomial of degree at most k.

Moreover, we require that the normal component $\boldsymbol{q}^h \cdot \boldsymbol{n}$ is continuous across each edge e of $\mathcal{E}_h$. This can be achieved by selecting the following element degrees of freedom:

(i) the moments of order up to k of $\boldsymbol{q}^h \cdot \boldsymbol{n}$ on the edges of T;

(ii) the moments of order up to $k - 1$ of $\boldsymbol{q}^h$ on T.

For the discretized temperature θ^h, we take piecewise polynomials of degree at most k (P_k element). The element degrees of freedom for the choice $k = 1$ are shown here:
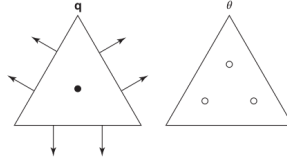
We now briefly consider the extension of some of the methods presented in the previous section to quadrilateral meshes. In this case, we define our approximating spaces on the reference element $\tilde{K} = [-1, 1]^2$ equipped with local coordinates (r, s). As far as the flux is concerned, the corresponding approximation space on each physical element K must be obtained through the use of a suitable transformation that preserves the normal component of vectorial functions. This is accomplished by the following (contravariant) Piola's transformation of vector fields. Suppose that

$$\boldsymbol{F} : \tilde{K} \to K; \qquad (x, y) = \boldsymbol{F}(r, s)$$

is an invertible map from $\tilde{K}$ onto K, with Jacobian matrix $\boldsymbol{J}(r, s)$. Given a vector field $\boldsymbol{q}(r, s)$ on $\hat{K}$, its Piola's transform $\mathcal{P}(\boldsymbol{q})(x, y)$ is the vector field on K defined by

$$\mathcal{P}(\boldsymbol{q})(x, y) := \frac{1}{|\boldsymbol{J}(r, s)|} \boldsymbol{J}(r, s) \boldsymbol{q}(r, s) \qquad (x, y) = \boldsymbol{F}(r, s)$$

Therefore if

$$\boldsymbol{Q}(\tilde{K}) = \text{Span}\{\boldsymbol{q}_i^h; \; i = 1, ... n_{el}\}$$

441

is an n_{el}-dimensional flux approximation space defined on the reference element $\tilde{K}$, the corresponding space on the physical element K will be

$$\boldsymbol{Q}(K) = \mathrm{Span}\{\mathcal{P}(\boldsymbol{q}_i^h);\ i = 1,...n_{el}\}$$

The $RT_{[0]} - P_0$ element. In the reference element K, we prescribe the approximated flux $\boldsymbol{q}^h$ as ($RT_{[0]}$ element)

$$\boldsymbol{q}^h = \left\{ \begin{array}{c} a + br \\ c + ds \end{array} \right\}, \qquad a, b, c \in \mathbb{R}$$

It is then easily seen from the equation above that the four values

$$\int_e \boldsymbol{q}^h \cdot \boldsymbol{n}\ ds \qquad \forall e \text{ edge of } \tilde{K}$$

can be chosen as a set of degrees of freedom. Moreover, $\mathrm{div}\boldsymbol{q}^h$ is constant in K, suggesting the choice of a constant approximated temperature θ^h in $\tilde{K}$ (P_0 element). The degrees of freedom for both $\boldsymbol{q}^h$ and θ^h are shown in this figure:
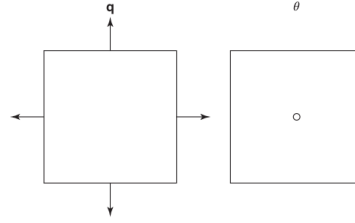
All the elements presented above have their three- dimensional counterpart. In this case, the normal component of the approximated flux $\boldsymbol{q}^h$ should not exhibit jumps across faces of adjacent elements.

In the following figure, we display the tetrahedral version of the $RT_0 - P_0$ element, consisting of a piecewise constant approximation for the temperature θ and of the following element approximating functions for $\boldsymbol{q}^h$ (Nedelec, 1980 [988]):

$$\boldsymbol{q}^h|_T = \boldsymbol{p}_0 + p_0 \left\{ \begin{array}{c} x \\ y \\ z \end{array} \right\} = \left\{ \begin{array}{c} a_0 + p_0 x \\ b_0 + p_0 y \\ c_0 + p_0 z \end{array} \right\} \qquad a_0, b_0, c_0 \in \mathbb{R}$$

(a)
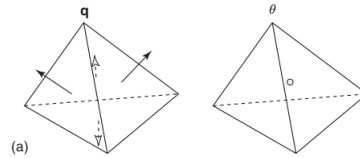
Therefore, in each tetrahedron T, the space for the approximated flux has dimension 4 and the degrees of freedom are precisely the values $\int_f \boldsymbol{q}^h \cdot \boldsymbol{n}\ d\sigma$ on each tetrahedron face f.

10.6.2   Degree 1 Raviart–Thomas on a triangle - Lagrange variant

https://defelement.com/elements/examples/triangle-raviart-thomas-lagrange-1.html

Taken from https://defelement.com/
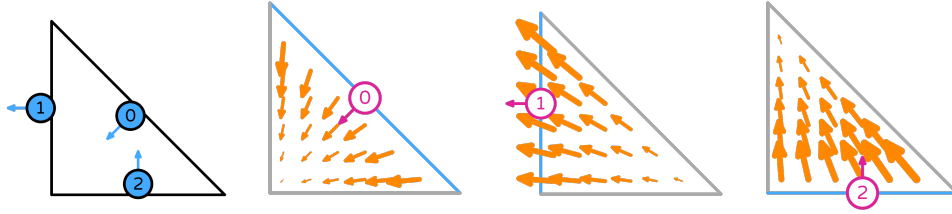
$\mathcal{V}$ is a finite dimensional polynomial space on the element of dimension 3 and is spanned by [404]

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ s \end{pmatrix}$$

Functionals and basis functions:

$$l_0 : \boldsymbol{v} \rightarrow \int_{e_0} \boldsymbol{v} \cdot (1)\hat{n}_0$$

$$l_1 : \boldsymbol{v} \rightarrow \int_{e_1} \boldsymbol{v} \cdot (1)\hat{n}_1$$

$$l_2 : \boldsymbol{v} \rightarrow \int_{e_2} \boldsymbol{v} \cdot (1)\hat{n}_2$$

where e_0 is the 0th edge and $\hat{n}_0$ is the normal to facet 0, where e_1 is the 1st edge and $\hat{n}_1$ is the normal to facet 1, where e_2 is the 2nd edge and $\hat{n}_2$ is the normal to facet 2.

We have then

$$\vec{\mathcal{N}}_0(r,s) = \begin{pmatrix} -r \\ -s \end{pmatrix} \qquad \vec{\mathcal{N}}_1(r,s) = \begin{pmatrix} r-1 \\ s \end{pmatrix} \qquad \vec{\mathcal{N}}_2(r,s) = \begin{pmatrix} -r \\ 1-s \end{pmatrix}$$

These DOFs are associated with edge 0,1,2 of the reference element.

10.6.3   Degree 2 Raviart–Thomas on a triangle - Lagrange variant
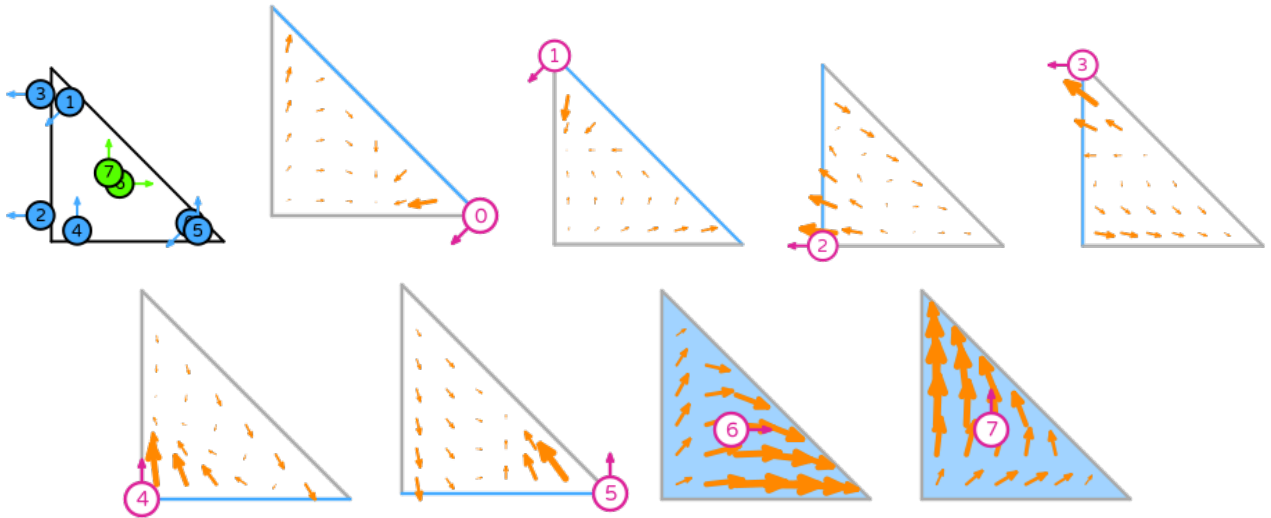
https://defelement.com/elements/examples/triangle-raviart-thomas-lagrange-2.html

Taken from https://defelement.com/

443

$\mathcal{V}$ is spanned by:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ r \end{pmatrix}, \begin{pmatrix} s \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ s \end{pmatrix}, \begin{pmatrix} r^2 \\ rs \end{pmatrix}, \begin{pmatrix} rs \\ s^2 \end{pmatrix}$$

This basis is also found in [404], example 14.5 (p.163).

We have then

$$\vec{\mathcal{N}_0}(r,s) = \begin{pmatrix} 4r(1-2r) \\ 2s(1-4r) \end{pmatrix}$$

$$\vec{\mathcal{N}_1}(r,s) = \begin{pmatrix} 2r(1-4s) \\ 4s(1-2s) \end{pmatrix}$$

$$\vec{\mathcal{N}_2}(r,s) = \begin{pmatrix} -8r^2 - 8rs + 12r + 6s - 4 \\ 2s(-4r - 4s + 3) \end{pmatrix}$$

$$\vec{\mathcal{N}_3}(r,s) = \begin{pmatrix} 8rs - 2r - 6s + 2 \\ 4s(2s - 1) \end{pmatrix}$$

$$\vec{\mathcal{N}_4}(r,s) = \begin{pmatrix} 2r(4r + 4s - 3) \\ 8rs - 6r + 8s^2 - 12s + 4 \end{pmatrix}$$

$$\vec{\mathcal{N}_5}(r,s) = \begin{pmatrix} 4r(1-2r) \\ -8rs + 6r + 2s - 2 \end{pmatrix}$$

$$\vec{\mathcal{N}_6}(r,s) = \begin{pmatrix} 8r(-2r - s + 2) \\ 8s(-2r - s + 1) \end{pmatrix}$$

$$\vec{\mathcal{N}_7}(r,s) = \begin{pmatrix} 8r(-r - 2s + 1) \\ 8s(-r - 2s + 2) \end{pmatrix}$$

These DOFs are associated with edge 0,1,2 of the reference element.

10.6.4 Degree 1 Raviart–Thomas on a quadrilateral - Lagrange variant
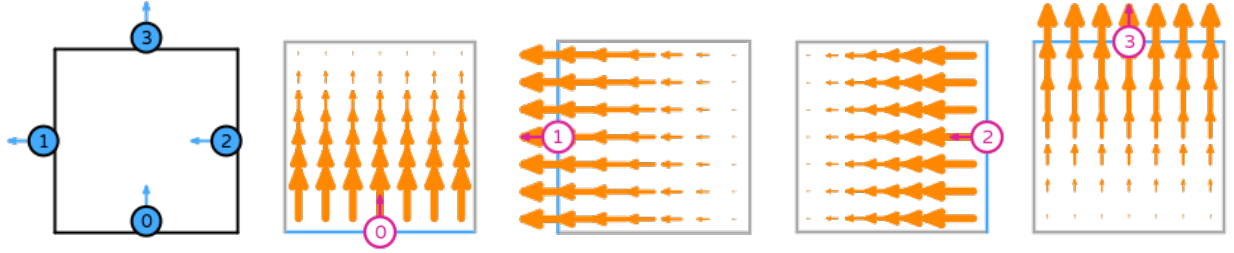
Taken from https://defelement.com/

$\mathcal{V}$ is a finite dimensional polynomial space on the element of dimension 4 and is spanned by

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ s \end{pmatrix}$$

Functionals and basis functions:

$$l_0 : \boldsymbol{v} \to \int_{e_0} \boldsymbol{v} \cdot (1)\hat{n}_0$$

$$l_1 : \boldsymbol{v} \to \int_{e_1} \boldsymbol{v} \cdot (1)\hat{n}_1$$

$$l_2 : \boldsymbol{v} \to \int_{e_2} \boldsymbol{v} \cdot (1)\hat{n}_2$$

444

$$l_3 : \boldsymbol{v} \to \int_{e_3} \boldsymbol{v} \cdot (1)\hat{n}_3$$

where e_0 is the 0th edge and $\hat{n}_0$ is the normal to facet 0, where e_1 is the 1st edge and $\hat{n}_1$ is the normal to facet 1, where e_2 is the 2nd edge and $\hat{n}_2$ is the normal to facet 2, where e_3 is the 4th edge and $\hat{n}_3$ is the normal to facet 3.

We have then

$$\vec{\mathcal{N}_0}(r,s) = \begin{pmatrix} 0 \\ 1-s \end{pmatrix} \qquad \vec{\mathcal{N}_1}(r,s) = \begin{pmatrix} r-1 \\ 0 \end{pmatrix} \qquad \vec{\mathcal{N}_2}(r,s) = \begin{pmatrix} -r \\ 0 \end{pmatrix} \qquad \vec{\mathcal{N}_3}(r,s) = \begin{pmatrix} 0 \\ s \end{pmatrix}$$

These DOFs are associated with edge 0,1,2 of the reference element.

10.6.5 Degree 1 Raviart–Thomas on a tetrahedron - Lagrange variant
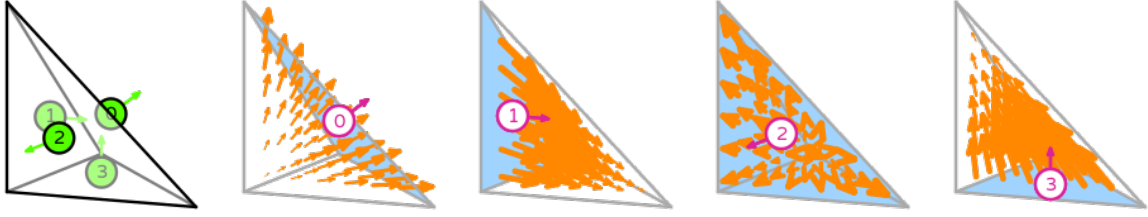
Taken from https://defelement.com/

$\mathcal{V}$ is a finite dimensional polynomial space on the element of dimension 4 and is spanned by

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ s \\ t \end{pmatrix}$$

$$\vec{\mathcal{N}_0}(r,s,t) = \begin{pmatrix} 2r \\ 2s \\ 2t \end{pmatrix} \quad \vec{\mathcal{N}_1}(r,s,t) = \begin{pmatrix} 2-2r \\ -2s \\ -2t \end{pmatrix} \quad \vec{\mathcal{N}_2}(r,s,t) = \begin{pmatrix} 2r \\ 2s-2 \\ 2t \end{pmatrix} \quad \vec{\mathcal{N}_3}(r,s,t) = \begin{pmatrix} -2r \\ -2s \\ 2-2t \end{pmatrix}$$

10.6.6 Degree 1 Raviart–Thomas on a hexahedron - Lagrange variant

Taken from https://defelement.com/
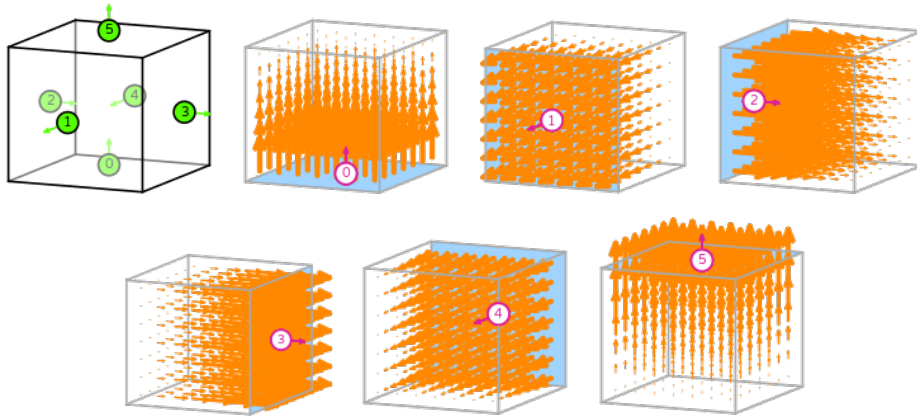
$\mathcal{V}$ is a finite dimensional polynomial space on the element of dimension 6 and is spanned by

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ s \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ t \end{pmatrix}.$$

445

$$\vec{\mathcal{N}}_0(r,s,t) = \begin{pmatrix} 0 \\ 0 \\ 1-t \end{pmatrix} \quad \vec{\mathcal{N}}_1(r,s,t) = \begin{pmatrix} 0 \\ s-1 \\ 0 \end{pmatrix} \quad \vec{\mathcal{N}}_2(r,s,t) = \begin{pmatrix} 1-r \\ 0 \\ 0 \end{pmatrix}$$

$$\vec{\mathcal{N}}_3(r,s,t) = \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix} \quad \vec{\mathcal{N}}_4(r,s,t) = \begin{pmatrix} 0 \\ -s \\ 0 \end{pmatrix} \quad \vec{\mathcal{N}}_5(r,s,t) = \begin{pmatrix} 0 \\ 0 \\ t \end{pmatrix}$$

What follows is mostly to be found in Weinberg and Sousedík [1446] (2019).

Let us consider a triangulation $\mathcal{T}_h$ of the domain Ω using triangles or quadrilaterals in 2D and tetrahedrons or hexahedrons in 3D. For simplicity, edges of elements in 2D will be called as faces, and faces in both 2D and 3D will be denoted by f_j. The discrete velocity space is defined as [404]

$$\mathcal{V}_h = \{\vec{v}|\vec{v}_{|K} \in RT_0(K), K \in \mathcal{T}_h \text{ and } \vec{\in} \mathcal{V}\}$$

with $\mathcal{V} \subset H(\text{div}, \Omega)$ and where RT_0 is the lowest-order Raviart-Thomas space on the element K. Let the basis (shape) functions for the velocity and pressure spaces be denoted $\mathcal{N}_i^\gamma$ and $\mathcal{N}_j^p$, respectively. The pressures are approximated by piecewise constant basis functions, and we denote by $\mathcal{Q}_h$ the discrete pressure space. For the velocity degrees of freedom we consider the average values of the normal components over the element faces. Specifically, the velocity degrees of freedom are defined as

$$\frac{1}{|f_j|} \int_{f_j} \vec{\mathcal{N}}_i^\gamma(\vec{x}_j) \cdot \vec{n}_j \, dl = \delta_{ij} \tag{10.41}$$

where δ_{ij} is the Kronecker delta, $\vec{n}_j$ is the unit outward normal of face j and $\vec{x}_j$ is its coordinate. In the calculations of the basis functions below, we use centers of the corresponding faces.

10.6.7 Triangle, RT0
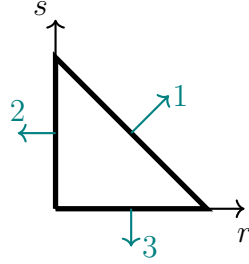
The basis functions in the reference triangle are defined as (see John reference earlier)

$$\vec{\mathcal{N}}_i(r,s) = \vec{a} + b \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} a_1 + br \\ a_2 + bs \end{pmatrix} \qquad i = 1, 2, 3$$

- face $j = 1$. Using Eq. (10.41) we can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1/2, 1/2)$, the normal is $\vec{n}_1 = (1/\sqrt{2}, 1/\sqrt{2})$ and $|f_1| = \sqrt{2}$):

$$
\begin{aligned}
\delta_{i1} = \frac{1}{|f_1|} \int_{f_1} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dl &= \frac{1}{\sqrt{2}} \int_{f_1} \vec{\mathcal{N}}_i^\gamma(1/2, 1/2) \cdot \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} dl \\
&= \frac{1}{\sqrt{2}} \int_{f_1} \begin{pmatrix} a_1 + b \cdot 1/2 \\ a_2 + b \cdot 1/2 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} dl \\
&= \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} (a_1 + a_2 + b) \int_{f_1} 1 \, dl \\
&= \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} (a_1 + a_2 + b) \sqrt{2} \\
&= \frac{1}{\sqrt{2}} (a_1 + a_2 + b) \tag{10.42}
\end{aligned}
$$

- face $j = 2$. Using Eq. (10.41) we can write for the second face (the middle of this face has coordinates $\vec{x}_2 = (0, 1/2)$, the normal is $\vec{n}_2 = (-1, 0)$ and $|f_2| = 1$):

447

$$\delta_{i2} = \frac{1}{|f_2|} \int_{f_2} \vec{\mathcal{N}}_i^\gamma(\vec{x}_2) \cdot \vec{n}_2 \, dl \;\; = \;\; \int_0^1 \vec{\mathcal{N}}_i^\gamma(0,1/2) \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix} ds$$

$$= \;\; \int_0^1 \begin{pmatrix} a_1 + b \cdot 0 \\ a_2 + b \cdot 1/2 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix} ds$$

$$= \;\; -a_1 \int_0^1 ds$$

$$= \;\; -a_1 \tag{10.43}$$

- face $j = 3$ Using Eq. (10.41) we can write for the third face (the middle of this face has coordinates $\vec{x}_3 = (1/2, 0)$, the normal is $\vec{n}_3 = (0, -1)$ and $|f_3| = 1$):

$$\delta_{i3} = \frac{1}{|f_3|} \int_{f_3} \vec{\mathcal{N}}_i^\gamma(\vec{x}_3) \cdot \vec{n}_3 \, dl \;\; = \;\; \int_0^1 \vec{\mathcal{N}}_i^\gamma(1/2,0) \cdot \begin{pmatrix} 0 \\ -1 \end{pmatrix} dr$$

$$= \;\; \int_0^1 \begin{pmatrix} a_1 + b \cdot 1/2 \\ a_2 + b \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -1 \end{pmatrix} dr$$

$$= \;\; -a_2 \int_0^1 dr$$

$$= \;\; -a_2 \tag{10.44}$$

We can now compute the three basis functions:

- Case $i = 1$: Using Eqs. (10.42),(10.43),(10.44) we can write

$$\frac{1}{\sqrt{2}}(a_1 + a_2 + b) \;\; = \;\; 1$$

$$-a_1 \;\; = \;\; 0$$

$$-a_2 \;\; = \;\; 0$$

which yields $a_1 = a_2 = 0$ and $b = \sqrt{2}$, i.e.

$$\vec{\mathcal{N}}_1^\gamma(r, s) = \sqrt{2} \begin{pmatrix} r \\ s \end{pmatrix}$$

- Case $i = 2$: Using Eqs. (10.42),(10.43),(10.44) we can write

$$\frac{1}{\sqrt{2}}(a_1 + a_2 + b) \;\; = \;\; 0$$

$$-a_1 \;\; = \;\; 1$$

$$-a_2 \;\; = \;\; 0$$

which yields $a_1 = -1$, $a_2 = 0$ and $b = -a_1 = 1$, i.e.

$$\vec{\mathcal{N}}_2^\gamma(r, s) = \begin{pmatrix} -1 + r \\ s \end{pmatrix}$$

- Case $i = 3$: Using Eqs. (10.42),(10.43),(10.44) we can write

$$\frac{1}{\sqrt{2}}(a_1 + a_2 + b) = 0$$
$$-a_1 = 0$$
$$-a_2 = 1$$

which yields $a_1 = 0$, $a_2 = -1$ and $b = -a_2 = 1$, i.e.

$$\vec{\mathcal{N}}_3^\gamma(r, s) = \begin{pmatrix} r \\ -1 + s \end{pmatrix}$$

In the end:

$$\vec{\mathcal{N}}_1^\gamma(r, s) = \sqrt{2}\begin{pmatrix} r \\ s \end{pmatrix} \qquad \vec{\mathcal{N}}_2^\gamma(r, s) = \begin{pmatrix} -1 + r \\ s \end{pmatrix} \qquad \vec{\mathcal{N}}_3^\gamma(r, s) = \begin{pmatrix} r \\ -1 + s \end{pmatrix}$$

10.6.8 Quadrilateral, RT0
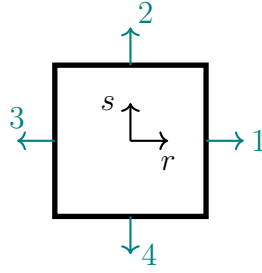
The basis functions are defined as (see literature section earlier on)

$$\vec{\mathcal{N}}_i(r, s) = \vec{a} + b\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} a_1 + b_1 r \\ a_2 + b_2 s \end{pmatrix} \qquad i = 1, 2, 3, 4$$

In this case all faces have the same length, i.e. $|f_j| = 2$.

- face $j = 1$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1, 0)$, the normal is $\vec{n}_1 = (1, 0)$):

$$
\begin{aligned}
\delta_{i1} = \frac{1}{|f_1|}\int_{f_1} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dl &= \frac{1}{2}\int_{f_1} \vec{\mathcal{N}}_i^\gamma(1, 0) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} dl \\
&= \frac{1}{2}\int_{f_1} \begin{pmatrix} a_1 + b_1 \cdot 1 \\ a_2 + b_2 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} dl \\
&= \frac{1}{2}\int_{-1}^{+1} (a_1 + b_1)ds \\
&= a_1 + b_1 \qquad\qquad (10.45)
\end{aligned}
$$

- face $j = 2$. We can write for the first face (the middle of this face has coordinates $\vec{x}_2 = (0, 1)$, the normal is $\vec{n}_2 = (0, 1)$):

449

$$\delta_{i2} = \frac{1}{|f_2|} \int_{f_2} \vec{\mathcal{N}}_i^\gamma(\vec{x}_2) \cdot \vec{n}_2 \, dl \;=\; \frac{1}{2} \int_{f_1} \vec{\mathcal{N}}_i^\gamma(0,1) \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} dl$$

$$= \frac{1}{2} \int_{f_1} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} dl$$

$$= \frac{1}{2} \int_{-1}^{+1} (a_2 + b_2) dr$$

$$= a_2 + b_2 \qquad\qquad (10.46)$$

- face $j = 3$. We can write for the first face (the middle of this face has coordinates $\vec{x}_3 = (-1, 0)$, the normal is $\vec{n}_3 = (-1, 0)$):

$$\delta_{i3} = \frac{1}{|f_3|} \int_{f_3} \vec{\mathcal{N}}_i^\gamma(\vec{x}_3) \cdot \vec{n}_3 \, dl \;=\; \frac{1}{2} \int_{f_1} \vec{\mathcal{N}}_i^\gamma(-1,0) \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix} dl$$

$$= \frac{1}{2} \int_{f_1} \begin{pmatrix} a_1 + b_1 \cdot -1 \\ a_2 + b_2 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix} dl$$

$$= -\frac{1}{2} \int_{-1}^{+1} (a_1 - b_1) ds$$

$$= -a_1 + b_1 \qquad\qquad (10.47)$$

- face $j = 4$. We can write for the first face (the middle of this face has coordinates $\vec{x}_4 = (0, -1)$, the normal is $\vec{n}_4 = (0, -1)$):

$$\delta_{i4} = \frac{1}{|f_4|} \int_{f_4} \vec{\mathcal{N}}_i^\gamma(\vec{x}_4) \cdot \vec{n}_4 \, dl \;=\; \frac{1}{2} \int_{f_1} \vec{\mathcal{N}}_i^\gamma(0,-1) \cdot \begin{pmatrix} 0 \\ -1 \end{pmatrix} dl$$

$$= \frac{1}{2} \int_{f_1} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot -1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -1 \end{pmatrix} dl$$

$$= -\frac{1}{2} \int_{-1}^{+1} (a_2 - b_2) dr$$

$$= -a_2 + b_2 \qquad\qquad (10.48)$$

- Case $i = 1$. We have to solve

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

which yields $a_1 = 1/2$, $b_1 = 1/2$, $a_2 = b_2 = 0$.

- Case $i = 2$. We have to solve

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

which yields $a_1 = 0$, $b_1 = 0$, $a_2 = 1/2$, $b_2 = 1/2$.

- Case $i = 3$. We have to solve

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

which $a_1 = -1/2$, $b_1 = 1/2$, $a_2 = 0$, $b_2 = 0$.

- Case $i = 4$. We have to solve

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

which yields $a_1 = 0$, $b_1 = 0$, $a_2 = -1/2$, $b_2 = 1/2$.

In the end we arrive at the following basis functions:

$$
\begin{aligned}
\vec{\mathcal{N}}_1^\gamma(r,s) &= \begin{pmatrix} \frac{1}{2}(1+r) \\ 0 \end{pmatrix} \\
\vec{\mathcal{N}}_2^\gamma(r,s) &= \begin{pmatrix} 0 \\ \frac{1}{2}(1+s) \end{pmatrix} \\
\vec{\mathcal{N}}_3^\gamma(r,s) &= \begin{pmatrix} \frac{1}{2}(-1+r) \\ 0 \end{pmatrix} \\
\vec{\mathcal{N}}_4^\gamma(r,s) &= \begin{pmatrix} 0 \\ \frac{1}{2}(-1+s) \end{pmatrix}
\end{aligned}
\tag{10.49}
$$

10.6.9   Tetrahedron, RT0

The basis functions in the reference triangle are defined as (see literature section earlier on)

$$\vec{\mathcal{N}}_i(r,s,t) = \vec{a} + b \begin{pmatrix} r \\ s \\ t \end{pmatrix} = \begin{pmatrix} a_1 + br \\ a_2 + bs \\ a_3 + bt \end{pmatrix} \qquad i = 1, 2, 3, 4$$

451
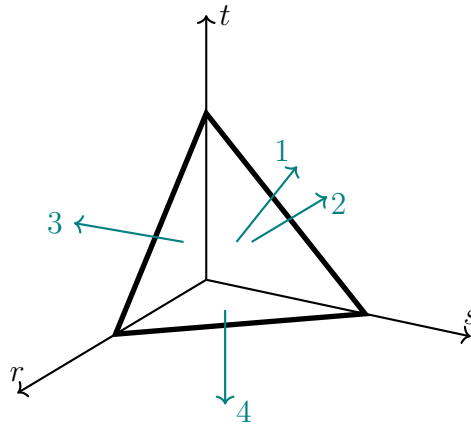
- face $j = 1$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1/3, 1/3, 1/3)$, the normal is $\vec{n}_1 = \frac{1}{\sqrt{3}}(1, 1, 1)$):

$$
\begin{aligned}
\delta_{i1} = \frac{1}{|f_1|} \iint_{f_1} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dS &= \frac{1}{|f_1|} \iint_{f_1} \vec{\mathcal{N}}_i^\gamma(1/3, 1/3, 1/3) \cdot \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} dS \\
&= \frac{1}{|f_1|} \iint_{f_1} \begin{pmatrix} a_1 + b \cdot 1/3 \\ a_2 + b \cdot 1/3 \\ a_3 + b \cdot 1/3 \end{pmatrix} \cdot \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} dS \\
&= \frac{1}{|f_1|} \iint_{f_1} (a_1 + a_2 + a_3 + b) dS \\
&= \frac{1}{\sqrt{3}} (a_1 + a_2 + a_3 + b) \quad\quad (10.50)
\end{aligned}
$$

- face $j = 2$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (0, 1/3, 1/3)$, the normal is $\vec{n}_1 = (-1, 0, 0)$):

$$
\begin{aligned}
\delta_{i2} = \frac{1}{|f_2|} \iint_{f_2} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dS &= \frac{1}{|f_2|} \iint_{f_2} \vec{\mathcal{N}}_i^\gamma(0, 1/3, 1/3) \cdot \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_2|} \iint_{f_2} \begin{pmatrix} a_1 + b \cdot 0 \\ a_2 + b \cdot 1/3 \\ a_3 + b \cdot 1/3 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_2|} \iint_{f_2} (-a_1) dS \\
&= -a_1 \quad\quad (10.51)
\end{aligned}
$$

- face $j = 3$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1/3, 0, 1/3)$, the normal is $\vec{n}_1 = (0, -1, 0)$):

$$
\begin{aligned}
\delta_{i3} = \frac{1}{|f_3|} \iint_{f_3} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dS &= \frac{1}{|f_3|} \iint_{f_3} \vec{\mathcal{N}}_i^\gamma(1/3, 0, 1/3) \cdot \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_3|} \iint_{f_3} \begin{pmatrix} a_1 + b \cdot 1/3 \\ a_2 + b \cdot 0 \\ a_3 + b \cdot 1/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_3|} \iint_{f_3} (-a_2) dS \\
&= -a_2 \quad\quad (10.52)
\end{aligned}
$$

- face $j = 4$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1/3, 1/3, 0)$, the normal is $\vec{n}_1 = (0, 0, -1)$):

$$\delta_{i4} = \frac{1}{|f_4|} \iint_{f_4} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dS \;\; = \;\; \frac{1}{|f_4|} \iint_{f_4} \vec{\mathcal{N}}_i^\gamma(1/3, 1/3, 0) \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} dS$$

$$= \;\; \frac{1}{|f_4|} \iint_{f_4} \begin{pmatrix} a_1 + b \cdot 1/3 \\ a_2 + b \cdot 1/3 \\ a_3 + b \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} dS$$

$$= \;\; \frac{1}{|f_4|} \iint_{f_4} (-a_3) dS$$

$$= \;\; -a_3 \tag{10.53}$$

- Case $i = 1$. We have to solve

$$\begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

which yields $a_1 = a_2 = a_3 = 0$ and $b = 1/\sqrt{3}$.

- Case $i = 2$. We have to solve

$$\begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

which yields $a_2 = a_3 = 0$, $a_1 = -1$ and $b = -a_1 = 1$.

- Case $i = 3$. We have to solve

$$\begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

which yields $a_1 = a_3 = 0$, $a_2 = -1$ and $b = -a_2 = 1$.

- Case $i = 4$. We have to solve

$$\begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

which yields $a_1 = a_2 = 0$, $a_3 = -1$ and $b = -a_3 = 1$.

In the end

$$\vec{\mathcal{N}}_1^\gamma(r,s) = \frac{1}{\sqrt{3}} \begin{pmatrix} r \\ s \\ t \end{pmatrix}$$

$$\vec{\mathcal{N}}_2^\gamma(r,s) = \begin{pmatrix} -1+r \\ s \\ t \end{pmatrix}$$

$$\vec{\mathcal{N}}_3^\gamma(r,s) = \begin{pmatrix} r \\ -1+s \\ t \end{pmatrix}$$

$$\vec{\mathcal{N}}_4^\gamma(r,s) = \begin{pmatrix} r \\ s \\ -1+t \end{pmatrix} \tag{10.54}$$

10.6.10 Hexahedron, RT0

The basis functions in the reference triangle are defined as (see literature section earlier on)

$$\vec{\mathcal{N}}_i(r,s,t) = \vec{a} + b \begin{pmatrix} r \\ s \\ t \end{pmatrix} = \begin{pmatrix} a_1 + b_1 r \\ a_2 + b_2 s \\ a_3 + b_3 t \end{pmatrix} \qquad i = 1,2,3,4,5,6$$
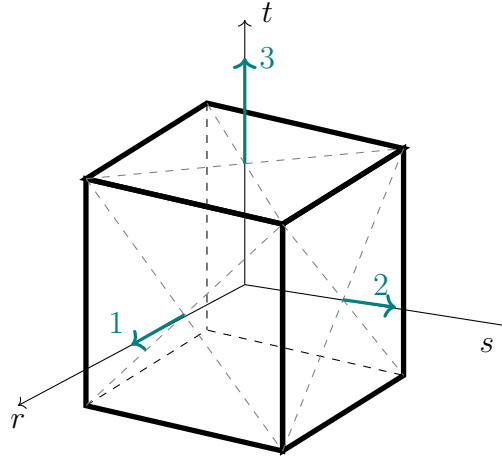
In this case all faces have the same length, i.e. $|f_j| = 4$.

- face $j = 1$. We can write for the first face (the middle of this face has coordinates $\vec{x}_1 = (1,0,0)$, the normal is $\vec{n}_1 = (1,0,0)$):

$$
\begin{aligned}
\delta_{i1} = \frac{1}{|f_1|} \iint_{f_1} \vec{\mathcal{N}}_i^\gamma(\vec{x}_1) \cdot \vec{n}_1 \, dS &= \frac{1}{|f_1|} \iint_{f_1} \vec{\mathcal{N}}_i^\gamma(1,0,0) \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_1|} \iint_{f_1} \begin{pmatrix} a_1 + b_1 \cdot 1 \\ a_2 + b_2 \cdot 0 \\ a_3 + b_3 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} dS \\
&= \frac{1}{|f_1|} \iint_{f_1} (a_1 + b_1) dS \\
&= a_1 + b_1 \tag{10.55}
\end{aligned}
$$

- face $j = 2$. We can write for the first face (the middle of this face has coordinates $\vec{x}_2 = (0, 1, 0)$, the normal is $\vec{n}_2 = (0, 1, 0)$):

$$\delta_{i2} = \frac{1}{|f_2|} \iint_{f_2} \vec{\mathcal{N}}_i^{\gamma}(\vec{x}_2) \cdot \vec{n}_2 \, dS = \frac{1}{|f_2|} \iint_{f_2} \vec{\mathcal{N}}_i^{\gamma}(0, 1, 0) \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} dS$$

$$= \frac{1}{|f_2|} \iint_{f_2} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot 1 \\ a_3 + b_3 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} dS$$

$$= \frac{1}{|f_2|} \iint_{f_2} (a_2 + b_2) dS$$

$$= a_2 + b_2 \tag{10.56}$$

- face $j = 3$. We can write for the first face (the middle of this face has coordinates $\vec{x}_3 = (0, 0, 1)$, the normal is $\vec{n}_3 = (0, 0, 1)$):

$$\delta_{i3} = \frac{1}{|f_3|} \iint_{f_3} \vec{\mathcal{N}}_i^{\gamma}(\vec{x}_3) \cdot \vec{n}_3 \, dS = \frac{1}{|f_3|} \iint_{f_3} \vec{\mathcal{N}}_i^{\gamma}(0, 0, 1) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} dS$$

$$= \frac{1}{|f_3|} \iint_{f_3} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot 0 \\ a_3 + b_3 \cdot 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} dS$$

$$= \frac{1}{|f_3|} \iint_{f_3} (a_3 + b_3) dS$$

$$= a_3 + b_3 \tag{10.57}$$

- face $j = 4$. We can write for the first face (the middle of this face has coordinates $\vec{x}_4 = (-1, 0, 0)$, the normal is $\vec{n}_4 = (-1, 0, 0)$):

$$\delta_{i4} = \frac{1}{|f_4|} \iint_{f_1} \vec{\mathcal{N}}_i^{\gamma}(\vec{x}_4) \cdot \vec{n}_4 \, dS = \frac{1}{|f_4|} \iint_{f_4} \vec{\mathcal{N}}_i^{\gamma}(-1, 0, 0) \cdot \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} dS$$

$$= \frac{1}{|f_4|} \iint_{f_4} \begin{pmatrix} a_1 + b_1 \cdot -1 \\ a_2 + b_2 \cdot 0 \\ a_3 + b_3 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} dS$$

$$= -\frac{1}{|f_4|} \iint_{f_4} (a_1 - b_1) dS$$

$$= -a_1 + b_1 \tag{10.58}$$

- face $j = 5$. We can write for the first face (the middle of this face has coordinates $\vec{x}_5 = (0, -1, 0)$, the normal is $\vec{n}_5 = (0, -1, 0)$):

$$\delta_{i5} = \frac{1}{|f_5|} \iint_{f_5} \vec{\mathcal{N}}_i^\gamma(\vec{x}_5) \cdot \vec{n}_5 \, dS = \frac{1}{|f_5|} \iint_{f_5} \vec{\mathcal{N}}_i^\gamma(0,-1,0) \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} dS$$

$$= \frac{1}{|f_5|} \iint_{f_5} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot -1 \\ a_3 + b_3 \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} dS$$

$$= -\frac{1}{|f_5|} \iint_{f_5} (a_2 - b_2) dS$$

$$= -a_2 + b_2 \tag{10.59}$$

- face $j = 6$. We can write for the first face (the middle of this face has coordinates $\vec{x}_6 = (0, 0, -1)$, the normal is $\vec{n}_6 = (0, 0, -1)$):

$$\delta_{i6} = \frac{1}{|f_6|} \iint_{f_6} \vec{\mathcal{N}}_i^\gamma(\vec{x}_6) \cdot \vec{n}_6 \, dS = \frac{1}{|f_6|} \iint_{f_6} \vec{\mathcal{N}}_i^\gamma(0,0,-1) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} dS$$

$$= \frac{1}{|f_6|} \iint_{f_6} \begin{pmatrix} a_1 + b_1 \cdot 0 \\ a_2 + b_2 \cdot 0 \\ a_3 + b_3 \cdot -1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} dS$$

$$= -\frac{1}{|f_6|} \iint_{f_6} (a_3 - b_3) dS$$

$$= -a_3 + b_3 \tag{10.60}$$

We cn now compute the basis functions:

- Case $i = 1$. We have to solve

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

which yields $a_1 = 1/2$, $b_1 = 1/2$, $a_2 = b_2 = a_3 = b_3 = 0$.

- Case $i = 2, 3, 4, 5, 6$. On the base of what is above and the derivations for the quadrilateral, I need not got through this.

The basis functions are then

$$
\begin{aligned}
\vec{\mathcal{N}}_1^\gamma(r,s,t) &= \begin{pmatrix} \frac{1}{2}(1+r) \\ 0 \\ 0 \end{pmatrix} \\[2ex]
\vec{\mathcal{N}}_2^\gamma(r,s,t) &= \begin{pmatrix} 0 \\ \frac{1}{2}(1+s) \\ 0 \end{pmatrix} \\[2ex]
\vec{\mathcal{N}}_3^\gamma(r,s,t) &= \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2}(1+t) \end{pmatrix} \\[2ex]
\vec{\mathcal{N}}_4^\gamma(r,s,t) &= \begin{pmatrix} \frac{1}{2}(-1+r) \\ 0 \\ 0 \end{pmatrix} \\[2ex]
\vec{\mathcal{N}}_5^\gamma(r,s,t) &= \begin{pmatrix} 0 \\ \frac{1}{2}(-1+s) \\ 0 \end{pmatrix} \\[2ex]
\vec{\mathcal{N}}_6^\gamma(r,s,t) &= \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2}(-1+t) \end{pmatrix}
\end{aligned}
\tag{10.61}
$$

🔖 Relevant Literature: Douglas N Arnold, Daniele Boffi, and Richard S Falk. "Quadrilateral H(div) finite elements". In: *SIAM Journal on Numerical Analysis* 42.6 (2005), pp. 2429–2451. DOI: 10.1137/S0036142903431924, Corinna Bahriawati and Carsten Carstensen. "Three MATLAB implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control". In: *Computational methods in applied mathematics* 5.4 (2005), pp. 333–361. DOI: 10.2478/cmam-2005-0016 , Theodore Weinberg and B Sousedík. "Fast implementation of mixed RT0 finite elements in MATLAB". in: *SIAM Undergraduate Research Online* 12 (2019). DOI: 10.1137/18S017430, P-A Raviart and Jean-Marie Thomas. "Primal hybrid finite element methods for 2nd order elliptic equations". In: *Mathematics of computation* 31.138 (1977), pp. 391–413. DOI: 10.1090/S0025-5718-1977-0431752-8 , Rolf Stenberg. "Some new families of finite elements for the Stokes equations". In: *Numerische Mathematik* 56 (1989), pp. 827–838. DOI: 10.1007/BF01405291, Douglas Arnold, Richard Falk, and Ragnar Winther. "Preconditioning in H(div) and applications". In: *Mathematics of computation* 66.219 (1997), pp. 957–984. DOI: xxx, Luca Bergamaschi, Stefano Mantica, and Fausto Saleri. "Mixed finite element approximation of Darcy's law in porous media". In: *Report CRS4 AppMath-94-20, CRS4. Cagliari, Italy* 21 (1994)

10.7 The penalty approach for viscous flow

In order to impose the incompressibility constraint, two widely used procedures are available, namely the Lagrange multiplier method and the penalty method [57, 637]. The latter allows for the elimination of the pressure variable from the momentum equation (resulting in a reduction of the matrix size).

Mathematical details on the origin and validity of the penalty approach applied to the Stokes problem can for instance be found in Cuvelier *et al.* [317], Reddy [1117] or Gunzburger [537].

The penalty formulation of the mass conservation equation is based on a relaxation of the incompressibility constraint and writes

$$\vec{\nabla} \cdot \vec{v} + \frac{p}{\lambda} = 0 \tag{10.62}$$

where λ is the penalty parameter, that can be interpreted (and has the same dimension) as a bulk viscosity. It is equivalent to say that the material is weakly compressible. It can be shown that if one chooses λ to be a sufficiently large number, the continuity equation $\vec{\nabla} \cdot \vec{v} = 0$ will be approximately satisfied in the finite element solution. The value of λ is often recommended to be 6 to 7 orders of magnitude larger than the shear viscosity [363, 641].

Equation (10.62) can be used to eliminate the pressure in the momentum equation so that the mass and momentum conservation equations fuse to become :

$$\vec{\nabla} \cdot (2\eta \dot{\varepsilon}(\vec{v})) + \lambda \vec{\nabla}(\vec{\nabla} \cdot \vec{v}) + \rho \vec{g} = \vec{0} \tag{10.63}$$

Malkus & Hughes (1978) [878] have established the equivalence for incompressible problems between the reduced integration of the penalty term and a mixed Finite Element approach if the pressure nodes coincide with the integration points of the reduced rule.

In the end, the elimination of the pressure unknown in the Stokes equations replaces the original saddle-point Stokes problem [77] by an elliptical problem, which leads to a symmetric positive definite (SPD) FEM matrix. This is the major benefit of the penalized approach over the full indefinite solver with the velocity-pressure variables. Indeed, the SPD character of the matrix lends itself to efficient solving stragegies and is less memory-demanding since it is sufficient to store only the upper half of the matrix including the diagonal [502].

The penalty approach for example is used in the SOPALE , DOUAR , ConMan, FANTOM and ELEFANT geodynamical codes.

Remark. *FEM codes relying on the penalty approach all rely on direct solvers, because as explained in Brezzi & Fortin [157]: "Using a penalty method is, for instance, almost impossible if an iterative method is used for the solution of the linear system, iterative methods being in general quite sensitive to the condition number of the matrix at hand."*

Since the penalty formulation is only valid for incompressible flows, then $\dot{\varepsilon}(\vec{v}) = \dot{\varepsilon}^d(\vec{v})$ so that the d superscript is omitted in what follows. We here focus on Cartesian coordinates only and because

the stress tensor is symmetric one can also rewrite it the following vector format:

$$
\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} = \begin{pmatrix} -p \\ -p \\ -p \\ 0 \\ 0 \\ 0 \end{pmatrix} + 2\eta \begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ \dot{\varepsilon}_{zz} \\ \dot{\varepsilon}_{xy} \\ \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yz} \end{pmatrix}
$$

$$
= \lambda \begin{pmatrix} \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} \\ \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} \\ \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} \\ 0 \\ 0 \\ 0 \end{pmatrix} + 2\eta \begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ \dot{\varepsilon}_{zz} \\ \dot{\varepsilon}_{xy} \\ \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yz} \end{pmatrix}
$$

$$
= \left[\lambda \underbrace{\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{\boldsymbol{K}} + \eta \underbrace{\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{C}} \right] \cdot \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial w}{\partial z} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \\ \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \end{pmatrix}
$$

Remember that

$$
\frac{\partial u^h}{\partial x} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial x} u_i \qquad \frac{\partial v^h}{\partial y} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial y} v_i \qquad \frac{\partial w^h}{\partial z} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial z} w_i
$$

and

$$
\frac{\partial u^h}{\partial y} + \frac{\partial v^h}{\partial x} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial y} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial x} v_i
$$

$$
\frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial x} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial z} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial x} w_i
$$

$$
\frac{\partial v^h}{\partial z} + \frac{\partial w^h}{\partial y} = \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial z} v_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i}{\partial y} w_i
$$

so that, since in $m_v = 8$ in 3D:

$$
\begin{pmatrix}
\frac{\partial u^h}{\partial x} \\[8pt]
\frac{\partial v^h}{\partial y} \\[8pt]
\frac{\partial w^h}{\partial z} \\[8pt]
\frac{\partial u^h}{\partial y} + \frac{\partial v^h}{\partial x} \\[8pt]
\frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial x} \\[8pt]
\frac{\partial v^h}{\partial z} + \frac{\partial w^h}{\partial y}
\end{pmatrix}
=
\underbrace{
\begin{pmatrix}
\frac{\partial \mathcal{N}_1}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_2}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_3}{\partial x} & 0 & 0 & \cdots & \frac{\partial \mathcal{N}_8}{\partial x} & 0 & 0 \\[6pt]
0 & \frac{\partial \mathcal{N}_1}{\partial y} & 0 & 0 & \frac{\partial \mathcal{N}_2}{\partial y} & 0 & 0 & \frac{\partial \mathcal{N}_3}{\partial y} & 0 & \cdots & 0 & \frac{\partial \mathcal{N}_8}{\partial y} & 0 \\[6pt]
0 & 0 & \frac{\partial \mathcal{N}_1}{\partial z} & 0 & 0 & \frac{\partial \mathcal{N}_2}{\partial z} & 0 & 0 & \frac{\partial \mathcal{N}_3}{\partial z} & \cdots & 0 & 0 & \frac{\partial \mathcal{N}_8}{\partial z} \\[6pt]
\frac{\partial \mathcal{N}_1}{\partial y} & \frac{\partial \mathcal{N}_1}{\partial x} & 0 & \frac{\partial \mathcal{N}_2}{\partial y} & \frac{\partial \mathcal{N}_2}{\partial x} & 0 & \frac{\partial \mathcal{N}_3}{\partial y} & \frac{\partial \mathcal{N}_3}{\partial x} & 0 & \cdots & \frac{\partial \mathcal{N}_8}{\partial y} & \frac{\partial \mathcal{N}_8}{\partial x} & 0 \\[6pt]
\frac{\partial \mathcal{N}_1}{\partial z} & 0 & \frac{\partial \mathcal{N}_1}{\partial x} & \frac{\partial \mathcal{N}_2}{\partial z} & 0 & \frac{\partial \mathcal{N}_2}{\partial x} & \frac{\partial \mathcal{N}_3}{\partial z} & 0 & \frac{\partial \mathcal{N}_3}{\partial x} & \cdots & \frac{\partial \mathcal{N}_8}{\partial z} & 0 & \frac{\partial \mathcal{N}_8}{\partial x} \\[6pt]
0 & \frac{\partial \mathcal{N}_1}{\partial z} & \frac{\partial \mathcal{N}_1}{\partial y} & 0 & \frac{\partial \mathcal{N}_2}{\partial z} & \frac{\partial \mathcal{N}_2}{\partial y} & 0 & \frac{\partial \mathcal{N}_3}{\partial z} & \frac{\partial \mathcal{N}_3}{\partial y} & \cdots & 0 & \frac{\partial \mathcal{N}_8}{\partial z} & \frac{\partial \mathcal{N}_8}{\partial y}
\end{pmatrix}
}_{\boldsymbol{B}(6\times 24)}
\cdot
\underbrace{
\begin{pmatrix}
u1 \\ v1 \\ w1 \\ u2 \\ v2 \\ w2 \\ u3 \\ v3 \\ w3 \\ \cdots \\ u8 \\ v8 \\ w8
\end{pmatrix}
}_{\vec{V}(24\times 1)}
$$

Finally,

$$
\vec{\sigma} =
\begin{pmatrix}
\sigma_{xx} \\
\sigma_{yy} \\
\sigma_{zz} \\
\sigma_{xy} \\
\sigma_{xz} \\
\sigma_{yz}
\end{pmatrix}
= (\lambda \boldsymbol{K} + \eta \boldsymbol{C}) \cdot \boldsymbol{B} \cdot \vec{V}
$$

We will now establish the weak form of the momentum conservation equation. We start again from

$$
\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{b} = \vec{0}
$$

For the $\mathcal{N}_i$'s 'regular enough', we can write:

$$
\int_{\Omega_e} \mathcal{N}_i \vec{\nabla} \cdot \boldsymbol{\sigma} \, dV + \int_{\Omega_e} \mathcal{N}_i \vec{b} \, dV = 0
$$

We can integrate by parts and drop the surface term[27]:

$$
\int_{\Omega_e} \vec{\nabla}\mathcal{N}_i \cdot \boldsymbol{\sigma} \, dV = \int_{\Omega_e} \mathcal{N}_i \vec{b} \, dV
$$

or,

$$
\int_{\Omega_e}
\begin{pmatrix}
\frac{\partial \mathcal{N}_i}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_i}{\partial y} & \frac{\partial \mathcal{N}_i}{\partial z} & 0 \\[6pt]
0 & \frac{\partial \mathcal{N}_i}{\partial y} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & 0 & \frac{\partial \mathcal{N}_i}{\partial z} \\[6pt]
0 & 0 & \frac{\partial \mathcal{N}_i}{\partial z} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & \frac{\partial \mathcal{N}_i}{\partial y}
\end{pmatrix}
\cdot
\begin{pmatrix}
\sigma_{xx} \\
\sigma_{yy} \\
\sigma_{zz} \\
\sigma_{xy} \\
\sigma_{xz} \\
\sigma_{yz}
\end{pmatrix}
dV = \int_{\Omega_e} \mathcal{N}_i \vec{b} \, dV
$$

---

[27]We will come back to this at a later stage

Let $i = 1, 2, 3, 4, \ldots 8$ and stack the resulting eight equations on top of one another.

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_i}{\partial y} & \frac{\partial \mathcal{N}_i}{\partial z} & 0 \\ 0 & \frac{\partial \mathcal{N}_i}{\partial y} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & 0 & \frac{\partial \mathcal{N}_i}{\partial z} \\ 0 & 0 & \frac{\partial \mathcal{N}_i}{\partial z} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & \frac{\partial \mathcal{N}_i}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} dV \;=\; \int_{\Omega_e} \mathcal{N}_1 \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} dV$$

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_i}{\partial y} & \frac{\partial \mathcal{N}_i}{\partial z} & 0 \\ 0 & \frac{\partial \mathcal{N}_i}{\partial y} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & 0 & \frac{\partial \mathcal{N}_i}{\partial z} \\ 0 & 0 & \frac{\partial \mathcal{N}_i}{\partial z} & 0 & \frac{\partial \mathcal{N}_i}{\partial x} & \frac{\partial \mathcal{N}_i}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} dV \;=\; \int_{\Omega_e} \mathcal{N}_2 \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} dV$$

$$\ldots$$

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial \mathcal{N}_8}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_8}{\partial y} & \frac{\partial \mathcal{N}_8}{\partial z} & 0 \\ 0 & \frac{\partial \mathcal{N}_8}{\partial y} & 0 & \frac{\partial \mathcal{N}_8}{\partial x} & 0 & \frac{\partial \mathcal{N}_8}{\partial z} \\ 0 & 0 & \frac{\partial \mathcal{N}_8}{\partial z} & 0 & \frac{\partial \mathcal{N}_8}{\partial x} & \frac{\partial \mathcal{N}_8}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} dV \;=\; \int_{\Omega_e} \mathcal{N}_8 \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} dV \quad (10.64)$$

We easily recognize $\boldsymbol{B}^T$ inside the integrals! Let us define

$$\vec{\mathcal{N}}_b^T = (\mathcal{N}_1 b_x, \mathcal{N}_1 b_y, \mathcal{N}_1 b_z \ldots \mathcal{N}_8 b_x, \mathcal{N}_8 b_y, \mathcal{N}_8 b_z)$$

then we can write

$$\int_{\Omega_e} \boldsymbol{B}^T \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} dV = \int_{\Omega_e} \vec{\mathcal{N}}_b \, dV$$

and finally:

$$\int_{\Omega_e} \boldsymbol{B}^T \cdot [\lambda \boldsymbol{K} + \eta \boldsymbol{C}] \cdot \boldsymbol{B} \cdot \vec{V} dV = \int_{\Omega_e} \vec{\mathcal{N}}_b \, dV$$

Since $\vec{V}$ is the vector of unknowns (i.e. the velocities at the corners), it does not depend on the x, y or z coordinates so it can be taken outside of the integral (remember $dV = dx \, dy \, dz$ here):

$$\underbrace{\left(\int_{\Omega_e} \boldsymbol{B}^T \cdot [\lambda \boldsymbol{K} + \eta \boldsymbol{C}] \cdot \boldsymbol{B} \, dV \right)}_{\boldsymbol{A}_{el}(24 \times 24)} \cdot \underbrace{\vec{V}}_{(24 \times 1)} = \underbrace{\int_{\Omega_e} \vec{\mathcal{N}}_b \, dV}_{\vec{B}_{el}(24 \times 1)}$$

or,

$$\left[\underbrace{\left(\int_{\Omega_e} \lambda \boldsymbol{B}^T \cdot \boldsymbol{K} \cdot \boldsymbol{B} \, dV \right)}_{\boldsymbol{A}_{el}^\lambda (24 \times 24)} + \underbrace{\left(\int_{\Omega_e} \eta \boldsymbol{B}^T \cdot \boldsymbol{C} \cdot \boldsymbol{B} \, dV \right)}_{\boldsymbol{A}_{el}^\eta (24 \times 24)} \right] \cdot \underbrace{\vec{V}}_{(24 \times 1)} = \underbrace{\int_{\Omega_e} \vec{\mathcal{N}}_b \, dV}_{\vec{B}_{el}(24 \times 1)}$$

Once the elemental matrix and rhs have been computed for an element its contribution is added to the global FEM matrix. The linear system is then solved and the velocity field at all nodes is obtained. From this velocity field the elemental pressure can be recovered by means of Eq. (10.62):

$$p = -\lambda \vec{\nabla} \cdot \vec{v}.$$
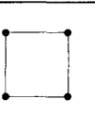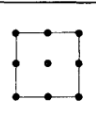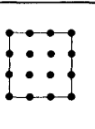
In two dimensions the equations above are very similar. Let us assume that the flow is taking place in the xy-plane and that the domain is infinite in the z-direction. Then $w = 0$ and $\partial_z \to 0$. From the 6 terms of the strain rate tensor only three remain: $\dot{\varepsilon}_{xx}$, $\dot{\varepsilon}_{yy}$ and $\dot{\varepsilon}_{xy}$. Then, since $m_v = 4$ (each element has 4 velocity nodes), we have

$$\boldsymbol{K} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{C} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Conversely the matrix $\boldsymbol{B}$ has size 3×8, etc ... Note that we will come across matrix $\boldsymbol{C}$ again when we solve the (non penalty-formulated) Stokes equations in the following sections.

As stated before the implementation is rather straightforward since only one FE matrix must be computed and assembled. However, there is one specific point which needs to be addressed: reduced integration.

To quote Hughes *et al.* (1979) [641]: *When a quadrature rule of lower order than the "standard" one is employed, this is called reduced integration. If all terms employ the same reduced integration, this is called uniform reduced integration; if reduced integration is used on some terms while standard integration is used on others, this is called selective reduced integration.*

shape functions	bilinear	biquadratic	bicubic
λ-term	1 point	2×2	3×3
μ-term	2×2	3×3	4×4

Selective Gauss-Legendre integration rules for 2-dimensional isoparametric Lagrange elements. Taken from Hughes *et al.* (1979) [641].

In the context of penalty-based codes, it has been shown [878] that it is crucial to resort to a selective reduced integration approach. The viscosity term $\boldsymbol{A}_{el}^{\eta}$ is integrated on 2^{ndim} points but the penalty term $\boldsymbol{A}_{el}^{\lambda}$ must be integrated on a single quadrature point. Finally, when the pressure is computed from the velocity via Eq. (10.62), the divergence term must also be computed at a single point in the middle of the element.

See STONE 01 for a concrete example of a 2D penalty-based Stokes solver.

🛠 Relevant Literature: Oden *et al.* [1008], Dhatt & Hubert[352].

10.8 The mixed FEM for viscous flow

10.8.1 In three dimensions

The FEM formulation of the Stokes equation is quite complex so we simplify things as much as possible for now by assuming the flow to be <u>incompressible</u>, <u>isoviscous</u> and <u>isothermal</u>.

The methodology to derive the discretised equations of the mixed system is quite similar to the one we have used in the case of the penalty formulation. The big difference comes from the fact that we are now solving for both velocity and pressure at the same time, and that we therefore must solve the mass and momentum conservation equations together. As before, velocity inside an element is given by

$$\vec{v}^h(\vec{r}) = \sum_{i=1}^{m_v} \mathcal{N}_i^{\gamma}(\vec{r}) \, \vec{v}_i \tag{10.65}$$

where $\mathcal{N}_i^{\gamma}$ are the polynomial basis functions for the velocity, and the summation runs over the m_v velocity nodes composing the element. A similar expression is used for pressure:

$$p^h(\vec{r}) = \sum_{i=1}^{m_p} \mathcal{N}_i^{p}(\vec{r}) \, p_i \tag{10.66}$$

Note that the velocity is a vector while pressure (and temperature) is a scalar. There are then $ndof_v = ndim$ velocity degrees of freedom per node and $ndof_p = 1$ pressure degrees of freedom. It is also very important to remember that the numbers of velocity nodes and pressure nodes for a given element are more often than not different and that velocity and pressure nodes need not be colocated. Indeed, unless so-called 'stabilised elements' are used, we have $m_v > m_p$, which means that the polynomial order of the velocity field is higher than the polynomial order of the pressure field (usually by value 1).

Other notations will be sometimes used for Eqs. (10.65) and (10.66):

$$u^h(\vec{r}) = \vec{\mathcal{N}}^{\gamma} \cdot \vec{u} \qquad v^h(\vec{r}) = \vec{\mathcal{N}}^{\gamma} \cdot \vec{v} \qquad w^h(\vec{r}) = \vec{\mathcal{N}}^{\gamma} \cdot \vec{w} \qquad p^h(\vec{r}) = \vec{\mathcal{N}}^{p} \cdot \vec{p} \tag{10.67}$$

where $\vec{v} = (u, v, w)$ and $\vec{\mathcal{N}}^{\gamma}$ is the vector containing all basis functions evaluated at location $\vec{r}$:

$$\vec{\mathcal{N}}^{v} = \left(\mathcal{N}_1^{\gamma}(\vec{r}), \mathcal{N}_2^{\gamma}(\vec{r}), \mathcal{N}_3^{\gamma}(\vec{r}), \dots \mathcal{N}_{m_v}^{\gamma}(\vec{r}) \right) \tag{10.68}$$

$$\vec{\mathcal{N}}^{p} = \left(\mathcal{N}_1^{p}(\vec{r}), \mathcal{N}_2^{p}(\vec{r}), \mathcal{N}_3^{p}(\vec{r}), \dots \mathcal{N}_{m_p}^{p}(\vec{r}) \right) \tag{10.69}$$

and with

$$\vec{u} = (u_1, u_2, u_3, \dots u_{m_v}) \tag{10.70}$$

$$\vec{v} = (v_1, v_2, v_3, \dots v_{m_v}) \tag{10.71}$$

$$\vec{w} = (w_1, w_2, w_3, \dots w_{m_v}) \tag{10.72}$$

$$\vec{p} = (p_1, p_2, p_3, \dots p_{m_p}) \tag{10.73}$$

We will now establish the weak form of the momentum conservation equation. We start again from

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{b} = \vec{0} \tag{10.74}$$

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{10.75}$$

For the $\mathcal{N}_i^\gamma$'s and $\mathcal{N}_i^p$ 'regular enough', we can write:

$$\int_{\Omega_e} \mathcal{N}_i^\gamma \vec{\nabla} \cdot \boldsymbol{\sigma} \, dV + \int_{\Omega_e} \mathcal{N}_i^\gamma \vec{b} \, dV \;=\; \vec{0} \tag{10.76}$$

$$\int_{\Omega_e} \mathcal{N}_i^p \vec{\nabla} \cdot \vec{v} \, dV \;=\; 0 \tag{10.77}$$

We can integrate by parts and drop the surface term[28]:

$$\int_{\Omega_e} \vec{\nabla} \mathcal{N}_i^\gamma \cdot \boldsymbol{\sigma} dV \;=\; \int_{\Omega_e} \mathcal{N}_i^\gamma \vec{b} \, dV \tag{10.78}$$

$$\int_{\Omega_e} \mathcal{N}_i^p \vec{\nabla} \cdot \vec{v} \, dV \;=\; 0 \tag{10.79}$$

or,

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial \mathcal{N}_i^\gamma}{\partial x} & 0 & 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial y} & \frac{\partial \mathcal{N}_i^\gamma}{\partial z} & 0 \\[2mm] 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial y} & 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial x} & 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial z} \\[2mm] 0 & 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial z} & 0 & \frac{\partial \mathcal{N}_i^\gamma}{\partial x} & \frac{\partial \mathcal{N}_i^\gamma}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} d\Omega = \int_{\Omega_e} \mathcal{N}_i^\gamma \vec{b} \, dV \tag{10.80}$$

The above equation can ultimately be written:

$$\int_{\Omega_e} \boldsymbol{B}^T \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} dV = \int_{\Omega_e} \vec{\mathcal{N}}_b \, dV \tag{10.81}$$

We have previously established that the strain rate vector $\vec{\dot{\varepsilon}}$ is:

$$\vec{\dot{\varepsilon}} = \begin{pmatrix} \frac{\partial u}{\partial x} \\[1mm] \frac{\partial v}{\partial y} \\[1mm] \frac{\partial w}{\partial z} \\[1mm] \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \\[1mm] \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \\[1mm] \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \end{pmatrix} = \begin{pmatrix} \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i \\[1mm] \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i \\[1mm] \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \\[1mm] \sum_i (\frac{\partial \mathcal{N}_i^\gamma}{\partial y} u_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial x} v_i) \\[1mm] \sum_i (\frac{\partial \mathcal{N}_i^\gamma}{\partial z} u_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial x} w_i) \\[1mm] \sum_i (\frac{\partial \mathcal{N}_i^\gamma}{\partial z} v_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial y} w_i) \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & 0 & 0 & \cdots & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial x} & 0 & 0 \\[1mm] 0 & \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & 0 & \cdots & 0 & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial y} & 0 \\[1mm] 0 & 0 & \frac{\partial \mathcal{N}_1^\gamma}{\partial z} & \cdots & 0 & 0 & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial z} \\[1mm] \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & 0 & \cdots & \frac{\partial N_{m_v}^\gamma}{\partial y} & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial x} & 0 \\[1mm] \frac{\partial \mathcal{N}_1^\gamma}{\partial z} & 0 & \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & \cdots & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial z} & 0 & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial x} \\[1mm] 0 & \frac{\partial \mathcal{N}_1^\gamma}{\partial z} & \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & \cdots & 0 & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial z} & \frac{\partial \mathcal{N}_{m_v}^\gamma}{\partial y} \end{pmatrix}}_{\boldsymbol{B}} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ w_1 \\ u_2 \\ v_2 \\ w_2 \\ u_3 \\ v_3 \\ \cdots \\ u_{m_v} \\ v_{m_v} \\ w_{m_v} \end{pmatrix}}_{\vec{\mathcal{V}}}$$
$$\tag{10.82}$$

or, $\vec{\dot{\varepsilon}} = \boldsymbol{B} \cdot \vec{\mathcal{V}}$ where $\boldsymbol{B}$ is the gradient matrix and $\vec{\mathcal{V}}$ is the vector of all velocity degrees of freedom for the element. The matrix $\boldsymbol{B}$ is then of size $6 \times (m_v \cdot ndof_v)$ and the vector $\vec{\mathcal{V}}$ is $m_v \cdot ndof_v$ long.

---

[28]We will come back to this at a later stage

we have

$$\sigma_{xx} = -p + 2\eta\dot{\varepsilon}_{xx}^d \tag{10.83}$$

$$\sigma_{yy} = -p + 2\eta\dot{\varepsilon}_{yy}^d \tag{10.84}$$

$$\sigma_{zz} = -p + 2\eta\dot{\varepsilon}_{zz}^d \tag{10.85}$$

$$\sigma_{xy} = 2\eta\dot{\varepsilon}_{xy}^d \tag{10.86}$$

$$\sigma_{xz} = 2\eta\dot{\varepsilon}_{xz}^d \tag{10.87}$$

$$\sigma_{yz} = 2\eta\dot{\varepsilon}_{yz}^d \tag{10.88}$$

Since we here only consider incompressible flow, we have $\dot{\boldsymbol{\varepsilon}}^d = \dot{\boldsymbol{\varepsilon}}$ so

$$\vec{\sigma} = - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} p + \boldsymbol{C} \cdot \vec{\dot{\varepsilon}} = - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \vec{N^p} \cdot \vec{P} + \boldsymbol{C} \cdot \boldsymbol{B} \cdot \vec{V} \tag{10.89}$$

with

$$\boldsymbol{C} = \eta \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \vec{\dot{\varepsilon}} = \begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ \dot{\varepsilon}_{zz} \\ 2\dot{\varepsilon}_{xy} \\ 2\dot{\varepsilon}_{xz} \\ 2\dot{\varepsilon}_{yz} \end{pmatrix} \tag{10.90}$$

Let us define matrix $\boldsymbol{\mathcal{N}}^p$ of size $6 \times m_p$:

$$\boldsymbol{\mathcal{N}}^p = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \vec{\mathcal{N}^p} = \begin{pmatrix} \vec{\mathcal{N}^p} \\ \vec{\mathcal{N}^p} \\ \vec{\mathcal{N}^p} \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{10.91}$$

so that

$$\vec{\sigma} = -\boldsymbol{\mathcal{N}}^p \cdot \vec{P} + \boldsymbol{C} \cdot \boldsymbol{B} \cdot \vec{V} \tag{10.92}$$

finally

$$\int_{\Omega_e} \boldsymbol{B}^T \cdot [-\boldsymbol{\mathcal{N}}^p \cdot \vec{P} + \boldsymbol{C} \cdot \boldsymbol{B} \cdot \vec{V}] \, d\Omega = \int_{\Omega_e} \boldsymbol{\mathcal{N}}_b \, d\Omega \tag{10.93}$$

or,

$$\underbrace{\left(-\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{\mathcal{N}}^p \, d\Omega \right)}_{\mathbb{G}} \cdot \vec{P} + \underbrace{\left(\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{C} \cdot \boldsymbol{B} \, d\Omega \right)}_{\mathbb{K}} \cdot \vec{V} = \underbrace{\int_{\Omega_e} \vec{\mathcal{N}}_b \, d\Omega}_{\vec{f}} \tag{10.94}$$

where the matrix $\mathbb{K}$ is of size $(m_v \cdot ndof_v \times m_v \cdot ndof_v)$, and matrix $\mathbb{G}$ is of size $(m_v \cdot ndof_v \times m_p \cdot ndof_p)$.

Turning now to the mass conservation equation:

$$
\begin{aligned}
\vec{0} &= \int_{\Omega_e} \vec{\mathcal{N}}^p \vec{\nabla} \cdot \vec{v} \, d\Omega \\
&= \int_{\Omega_e} \vec{\mathcal{N}}^p \sum_{i=1}^{m_v} \left(\frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \right) d\Omega \\
&= \int_{\Omega_e} \begin{pmatrix} \mathcal{N}_1^p \left(\sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \right) \\ \mathcal{N}_2^p \left(\sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \right) \\ \mathcal{N}_3^p \left(\sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \right) \\ \cdots \\ \mathcal{N}_{m_p}^p \left(\sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i + \sum_{i=1}^{m_v} \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \right) \end{pmatrix} dV \\
&= \int_{\Omega_e} \begin{pmatrix} \mathcal{N}_1^p & \mathcal{N}_1^p & \mathcal{N}_1^p & 0 & 0 & 0 \\ \mathcal{N}_2^p & \mathcal{N}_2^p & \mathcal{N}_2^p & 0 & 0 & 0 \\ \mathcal{N}_3^p & \mathcal{N}_3^p & \mathcal{N}_3^p & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{N}_{m_p}^p & \mathcal{N}_{m_p}^p & \mathcal{N}_{m_p}^p & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial x} u_i \\ \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial y} v_i \\ \sum_i \frac{\partial \mathcal{N}_i^\gamma}{\partial z} w_i \\ \sum_i \left(\frac{\partial \mathcal{N}_i^\gamma}{\partial y} u_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial x} v_i \right) \\ \sum_i \left(\frac{\partial \mathcal{N}_i^\gamma}{\partial z} u_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial x} w_i \right) \\ \sum_i \left(\frac{\partial \mathcal{N}_i^\gamma}{\partial z} v_i + \frac{\partial \mathcal{N}_i^\gamma}{\partial y} w_i \right) \end{pmatrix} dV \\
&= \int_{\Omega_e} \underbrace{\begin{pmatrix} \mathcal{N}_1^p & \mathcal{N}_1^p & \mathcal{N}_1^p & 0 & 0 & 0 \\ \mathcal{N}_2^p & \mathcal{N}_2^p & \mathcal{N}_2^p & 0 & 0 & 0 \\ \mathcal{N}_3^p & \mathcal{N}_3^p & \mathcal{N}_3^p & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{N}_{m_p}^p & \mathcal{N}_{m_p}^p & \mathcal{N}_{m_p}^p & 0 & 0 & 0 \end{pmatrix}}_{(\boldsymbol{\mathcal{N}}^p)^T} \cdot \vec{\varepsilon} \, dV \\
&= \left(\int (\boldsymbol{\mathcal{N}}^p)^T \cdot \boldsymbol{B} \, dV \right) \cdot \vec{V} \\
&= -\mathbb{G}_e^T \cdot \vec{V} \tag{10.95}
\end{aligned}
$$

Note that it is common to actually start from $-\vec{\nabla} \cdot \vec{v} = 0$ (see Eq.(3) in [897]) so as to arrive at $\mathbb{G}_e^T \cdot \vec{V} = \vec{0}$

Ultimately we obtain the following system for each element:

$$
\begin{pmatrix} \mathbb{K}_e & \mathbb{G}_e \\ -\mathbb{G}_e^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{V} \\ \vec{P} \end{pmatrix} = \begin{pmatrix} \vec{f}_e \\ 0 \end{pmatrix}
$$

Such a matrix is then generated for each element and then must me assembled into the global F.E. matrix. Note that in this case the elemental Stokes matrix is antisymmetric. One can also define

the following symmetric modified Stokes matrix:

$$\begin{pmatrix} \mathbb{K}_e & \mathbb{G}_e \\ \mathbb{G}_e^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{V} \\ \vec{P} \end{pmatrix} = \begin{pmatrix} \vec{f_e} \\ 0 \end{pmatrix} \tag{10.96}$$

This matrix is symmetric, but indefinite. It is non-singular if $ker(\mathbb{G}^T) = 0$, which is the case if the compatibility condition holds.

CHECK: Matrix $\mathbb{K}$ is the viscosity matrix. Its size is $(ndof_v * N_v) \times (ndof_v * N_v)$ where $ndof_v$ is the number of velocity degrees of freedom per node (typically 1,2 or 3) and N_v is the number of velocity nodes. The size of matrix $\mathbb{G}$ is $(ndof_v * N_v) \times (ndof_p * N_p)$ where $ndof_p (= 1)$ is the number of velocity degrees of freedom per node and N_p is the number of pressure nodes. Conversely, the size of matrix $\mathbb{G}^T$ is $(ndof_p * N_p) \times (ndof_v * N_v)$. The size of the global FE matrix is $N = ndof_v * N_v + ndof_p * N_p$. Note that matrix $\mathbb{K}$ is analogous to a discrete Laplacian operator, matrix $\mathbb{G}$ to a discrete gradient operator, and matrix $\mathbb{G}^T$ to a discrete divergence operator.

On the physical dimensions of the Stokes matrix blocks

We start from the Stokes equations:

$$-\vec{\nabla} p + \vec{\nabla} \cdot (2\eta \dot{\varepsilon}) + \rho \vec{g} = \vec{0} \tag{10.97}$$

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{10.98}$$

We have $[p] = ML^{-1}T^{-2}$, $[\vec{\nabla}] = L^{-1}$, so the dimensions of the terms in the first equation are: $ML^{-2}T^{-2}$. The blocks $\mathbb{K}$ and $\mathbb{G}$ stem from the weak form which is obtained by multiplying the strong form equations by the (dimensionless) basis functions and integrating over the 3D domain, so that it follows that

$$[\mathbb{K} \cdot \vec{\mathcal{V}}] = [\mathbb{G} \cdot \vec{\mathcal{P}}] = [\vec{f}] = (ML^{-2}T^{-2}) \cdot L^3 = MLT^{-2}$$

We can then easily deduce:

$$[\mathbb{K}] = MT^{-1} \qquad [\mathbb{G}] = L^2$$

Turning to the mass conservation equation, we have $[\vec{\nabla} \cdot \vec{v}] = L^{-1}LT^{-1} = T^{-1}$, which yields the discretised weak form $\mathbb{G} \cdot \vec{\mathcal{V}} = 0$ so that $[\mathbb{G} \cdot \vec{\mathcal{V}}] = L^3 T^{-1}$ and we of course recover $[\mathbb{G}] = L^2$.

If we wanted both equations to have the same dimensions, we would need to multiply the second one by a characteristic quantity which dimension is $ML^{-2}T^{-1}$, i.e. for example η/L (since $[\eta] = ML^{-1}T^{-1}$). This is indeed what we end up doing in practice, see Section 10.8.4.

On elemental level mass balance

Note that in what is above no assumption has been made about whether the pressure basis functions are continuous or discontinuous from one element to another.

Indeed, as mentioned in Gresho & Sani [517], since the weak formulation of the momentum equation involves integration by parts of $\vec{\nabla} p$, the resulting weak form contains no derivatives of pressure. This introduces the possibility of approximating it by functions (piecewise polynomials, of course) that are not C^0-continuous, and indeed this has been done and is quite popular/useful (e.g. P_0 or P_{-1}).

It is then worth noting that *only* discontinuous pressure elements assure an element-level mass balance [517]: if for instance $\mathcal{N}_i^p$ is piecewise-constant on element e (of value 1), the elemental weak form of the mass conservation equation is

$$\int_{\Omega_e} N_i^p \vec{\nabla} \cdot \vec{v} = \int_{\Omega_e} \vec{\nabla} \cdot \vec{v} = \int_{\Gamma_e} \vec{n} \cdot \vec{v} = 0$$

One potentially unwelcome consequence of using discontinuous pressure elements is that they do not possess uniquely defined pressure on the element boundaries; they are dual valued there, and often multi-valued at certain velocity nodes.

On the C matrix

The relationship between deviatoric stress and deviatoric strain rate tensor is

$$\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}^d \tag{10.99}$$

$$= 2\eta\left(\dot{\boldsymbol{\varepsilon}} - \frac{1}{3}(\vec{\nabla}\cdot\vec{v})\mathbf{1}\right) \tag{10.100}$$

$$= 2\eta\left[\begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} & \dot{\varepsilon}_{yz} \\ \dot{\varepsilon}_{zx} & \dot{\varepsilon}_{zy} & \dot{\varepsilon}_{zz} \end{pmatrix} - \frac{1}{3}(\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz})\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right] \tag{10.101}$$

$$= \frac{2}{3}\eta\begin{pmatrix} 2\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} - \dot{\varepsilon}_{zz} & 3\dot{\varepsilon}_{xy} & 3\dot{\varepsilon}_{xz} \\ 3\dot{\varepsilon}_{yx} & -\dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{yy} & 3\dot{\varepsilon}_{yz} \\ 3\dot{\varepsilon}_{zx} & 3\dot{\varepsilon}_{zy} & -\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{zz} \end{pmatrix} \tag{10.102}$$

so that

$$\vec{\tau} = \frac{2}{3}\eta\begin{pmatrix} 2\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} - \dot{\varepsilon}_{zz} \\ -\dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{yy} \\ -\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy} + 2\dot{\varepsilon}_{zz} \\ 3\dot{\varepsilon}_{xy} \\ 3\dot{\varepsilon}_{xz} \\ 3\dot{\varepsilon}_{yz} \end{pmatrix} = \frac{\eta}{3}\underbrace{\begin{pmatrix} 4 & -2 & -2 & 0 & 0 & 0 \\ -2 & 4 & -2 & 0 & 0 & 0 \\ -2 & -2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}}_{\boldsymbol{C}^d}\cdot\begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ \dot{\varepsilon}_{zz} \\ 2\dot{\varepsilon}_{xy} \\ 2\dot{\varepsilon}_{xz} \\ 2\dot{\varepsilon}_{yz} \end{pmatrix} = \boldsymbol{C}^d\cdot\vec{\dot{\varepsilon}} \tag{10.103}$$

which is identical to the one in the Appendix A of Schmalholz (2008) [1194]. In two dimensions, we have

$$\vec{\tau} = \frac{1}{3}\eta\underbrace{\begin{pmatrix} 4 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 3 \end{pmatrix}}_{\boldsymbol{C}^d}\cdot$$

see for instance Andres-Martinez *et al.* (2015) [24].

In the case where we assume incompressible flow from the beginning, i.e. $\dot{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}}^d$, then

$$\vec{\tau} = \eta\underbrace{\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{C}}\cdot\begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ \dot{\varepsilon}_{zz} \\ 2\dot{\varepsilon}_{xy} \\ 2\dot{\varepsilon}_{xz} \\ 2\dot{\varepsilon}_{yz} \end{pmatrix} = \boldsymbol{C}\cdot\vec{\dot{\varepsilon}} \tag{10.104}$$

A slightly different formulation

The momentum conservation equation can be written as follows:

$$\vec{\nabla}\cdot(2\eta\dot{\boldsymbol{\varepsilon}}(\vec{v})) - \vec{\nabla}p + \vec{b} = \vec{0}$$

When the viscosity η is constant and the flow is incompressible this equation becomes

$$\eta \Delta \vec{v} - \vec{\nabla} p + \vec{b} = \vec{0}$$

In this case the matrix $\boldsymbol{B}$ takes a different form (See Donea & Huerta [363, Eq. 6.24]) and one should be aware that this can have consequences for the Neumann boundary conditions.

In Burstedde *et al.* (2009) [202] the authors state that when the Laplacian formulation is used it has the computational advantage that the velocity components are coupled only through the incompressibility condition. While the two formulations are equivalent only for constant viscosity, they state that they employ the Laplacian approach formulation as a preconditioner for the viscous term.

Concretely, we apply the same method as above, i.e. we reorganise the terms of the velocity gradient tensor in a vector:

$$\vec{\nabla}\vec{v} \rightarrow \begin{pmatrix} \partial_x u \\ \partial_y u \\ \partial_z u \\ \partial_x v \\ \partial_y v \\ \partial_z v \\ \partial_x w \\ \partial_y w \\ \partial_z w \end{pmatrix} = \begin{pmatrix} \sum_i \partial_x \mathcal{N}_i u_i \\ \sum_i \partial_y \mathcal{N}_i u_i \\ \sum_i \partial_z \mathcal{N}_i u_i \\ \sum_i \partial_x \mathcal{N}_i v_i \\ \sum_i \partial_y \mathcal{N}_i v_i \\ \sum_i \partial_z \mathcal{N}_i v_i \\ \sum_i \partial_x \mathcal{N}_i w_i \\ \sum_i \partial_y \mathcal{N}_i w_i \\ \sum_i \partial_z \mathcal{N}_i w_i \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} \partial_x \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_x \mathcal{N}_2^{\gamma} & 0 & 0 & \cdots & \partial_x \mathcal{N}_{m_v}^{\gamma} & 0 & 0 \\ \partial_y \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_y \mathcal{N}_2^{\gamma} & 0 & 0 & \cdots & \partial_y \mathcal{N}_{m_v}^{\gamma} & 0 & 0 \\ \partial_z \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_z \mathcal{N}_2^{\gamma} & 0 & 0 & \cdots & \partial_z \mathcal{N}_{m_v}^{\gamma} & 0 & 0 \\ 0 & \partial_x \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_x \mathcal{N}_2^{\gamma} & 0 & \cdots & 0 & \partial_x \mathcal{N}_{m_v}^{\gamma} & 0 \\ 0 & \partial_y \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_y \mathcal{N}_2^{\gamma} & 0 & \cdots & 0 & \partial_y \mathcal{N}_{m_v}^{\gamma} & 0 \\ 0 & \partial_z \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_z \mathcal{N}_2^{\gamma} & 0 & \cdots & 0 & \partial_z \mathcal{N}_{m_v}^{\gamma} & 0 \\ 0 & 0 & \partial_x \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_x \mathcal{N}_2^{\gamma} & \cdots & 0 & 0 & \partial_x \mathcal{N}_{m_v}^{\gamma} \\ 0 & 0 & \partial_y \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_y \mathcal{N}_2^{\gamma} & \cdots & 0 & 0 & \partial_y \mathcal{N}_{m_v}^{\gamma} \\ 0 & 0 & \partial_z \mathcal{N}_1^{\gamma} & 0 & 0 & \partial_z \mathcal{N}_2^{\gamma} & \cdots & 0 & 0 & \partial_z \mathcal{N}_{m_v}^{\gamma} \end{pmatrix}}_{\boldsymbol{B}} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ w_1 \\ u_2 \\ v_2 \\ w_2 \\ u_3 \\ v_3 \\ \cdots \\ u_{m_v} \\ v_{m_v} \\ w_{m_v} \end{pmatrix}}_{\vec{V}}$$

and in two dimensions:

$$\vec{\nabla}\vec{v} \rightarrow \begin{pmatrix} \partial_x u \\ \partial_y u \\ \partial_x v \\ \partial_y v \end{pmatrix} = \begin{pmatrix} \sum_i \partial_x \mathcal{N}_i u_i \\ \sum_i \partial_y \mathcal{N}_i u_i \\ \sum_i \partial_x \mathcal{N}_i v_i \\ \sum_i \partial_y \mathcal{N}_i v_i \end{pmatrix} = \underbrace{\begin{pmatrix} \partial_x \mathcal{N}_1^{\gamma} & 0 & \partial_x \mathcal{N}_2^{\gamma} & 0 & \cdots & \partial_x \mathcal{N}_i^{\gamma} m_v & 0 \\ \partial_y \mathcal{N}_1^{\gamma} & 0 & \partial_y \mathcal{N}_2^{\gamma} & 0 & \cdots & \partial_y \mathcal{N}_i^{\gamma} m_v & 0 \\ 0 & \partial_x \mathcal{N}_1^{\gamma} & 0 & \partial_x \mathcal{N}_2^{\gamma} & \cdots & 0 & \partial_x \mathcal{N}_i^{\gamma} m_v \\ 0 & \partial_y \mathcal{N}_1^{\gamma} & 0 & \partial_y \mathcal{N}_2^{\gamma} & \cdots & 0 & \partial_y \mathcal{N}_i^{\gamma} m_v \end{pmatrix}}_{\boldsymbol{B}} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ \cdots \\ u_{m_v} \\ v_{m_v} \end{pmatrix}}_{\vec{V}}$$

If such a formulation is used, it makes more sense to actually group the unknowns as follows:

$$\vec{\mathcal{V}} = (u_1, \ldots, u_{m_v}, v_1, \ldots, v_{m_v}, w_1, \ldots, w_{m_v})$$

We start from

$$\eta \Delta \vec{v} - \vec{\nabla} p + \rho \vec{g} = \vec{0}$$

In 2D Cartesian coordinates this becomes:

$$\eta \Delta u - \partial_x p + \rho g_x \quad = \quad 0 \tag{10.105}$$
$$\eta \Delta v - \partial_y p + \rho g_y \quad = \quad 0 \tag{10.106}$$

or,

$$\eta \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - \partial_x p + \rho g_x \quad = \quad 0 \tag{10.107}$$
$$\eta \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) - \partial_y p + \rho g_y \quad = \quad 0 \tag{10.108}$$

Assuming that we prescribe the normal velocity on all sides (i.e. no Neumann boundary conditions), we can establish the weak form of these equations:

$$\underbrace{\left(\int_\Omega \eta \left(\frac{\partial \vec{\mathcal{N}}^v}{\partial x} \frac{\partial \vec{\mathcal{N}}^v}{\partial x} + \frac{\partial \vec{\mathcal{N}}^v}{\partial y} \frac{\partial \vec{\mathcal{N}}^v}{\partial y} \right) dV \right)}_{\mathbb{N}} \cdot \vec{\mathcal{V}}_x + \underbrace{\left(- \int_\Omega \frac{\partial \vec{\mathcal{N}}^v}{\partial x} \vec{\mathcal{N}}^p \, dV \right)}_{G_x} \cdot \vec{\mathcal{P}} \quad = \quad \underbrace{\int_\Omega \vec{\mathcal{N}}^v \rho g_x \, dV}_{\vec{f}_x} \tag{10.109}$$

$$\underbrace{\left(\int_\Omega \eta \left(\frac{\partial \vec{\mathcal{N}}^v}{\partial x} \frac{\partial \vec{\mathcal{N}}^v}{\partial x} + \frac{\partial \vec{\mathcal{N}}^v}{\partial y} \frac{\partial \vec{\mathcal{N}}^v}{\partial y} \right) dV \right)}_{\mathbb{N}} \cdot \vec{\mathcal{V}}_y + \underbrace{\left(- \int_\Omega \frac{\partial \vec{\mathcal{N}}^v}{\partial y} \vec{\mathcal{N}}^p \, dV \right)}_{G_y} \cdot \vec{\mathcal{P}} \quad = \quad \underbrace{\int_\Omega \vec{\mathcal{N}}^v \rho g_y \, dV}_{\vec{f}_y} \tag{10.110}$$

Turning now to the continuity equation

$$-\vec{\nabla} \cdot \vec{v} = 0$$

or,

$$-\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0$$

Its weak form then is

$$\underbrace{\left(- \int_\Omega \vec{\mathcal{N}}^p \frac{\partial \vec{\mathcal{N}}^v}{\partial x} dV \right)}_{G_x^T} \cdot \vec{\mathcal{V}}_x + \underbrace{\left(- \int_\Omega \vec{\mathcal{N}}^p \frac{\partial \vec{\mathcal{N}}^v}{\partial y} dV \right)}_{G_y^T} \cdot \vec{\mathcal{V}}_y = 0$$

In the end:

$$\begin{pmatrix} \mathbb{N} & 0 & \mathbb{G}_x \\ 0 & \mathbb{N} & \mathbb{G}_y \\ \mathbb{G}_x & \mathbb{G}_y & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}}_x \\ \vec{\mathcal{V}}_y \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f}_x \\ \vec{f}_y \\ 0 \end{pmatrix}$$

This approach is implemented in STONE 48.

On the 'forgotten' surface terms

FINISH write

10.8.2 Revisiting the penalty method

We have just seen that the discretised Stokes equation yield the following saddle point system:

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{0} \end{pmatrix}
$$

One can perturb the continuity equation by a term $\mathbb{C}_\epsilon = \epsilon \mathbb{M}_p$ where $\mathbb{M}_p$ is the pressure mass matrix. This yields

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C}_\epsilon \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{0} \end{pmatrix}
$$

or,

$$
\vec{\mathcal{P}} = \frac{1}{\epsilon} \mathbb{M}_p^{-1} \cdot \mathbb{G}^T \cdot \vec{\mathcal{V}}
$$

Substituting pressure in the first equation yields:

$$
(\mathbb{K} + \frac{1}{\epsilon} \mathbb{G} \cdot \mathbb{M}_p^{-1} \cdot \mathbb{G}^T) \cdot \vec{\mathcal{V}} = \vec{f} \tag{10.111}
$$

If we want to solve these equations, it is necessary that the matrix $\mathbb{M}_p^{-1}$ can be computed easily. This is for example the case if $\mathbb{M}_p$ is a lumped mass matrix (often done for Taylor-Hood elements). When discontinuous pressure elements are used, $\mathbb{M}_p$ is in a block diagonal matrix, i.e. a diagonal matrix consisting of small matrices as diagonal elements. One can easily verify that these small matrices have the size of the number of pressure unknowns per element. Note that this is all carried out at the elemental level.

10.8.3 A much more compact derivation of the Stokes matrix blocks

What follows is inspired by chapter 6 of Donea and Huerta [363]. One can easily show that the weak form of the Stokes system can be written

$$
\int_\Omega \vec{\nabla}\vec{\omega} : \boldsymbol{\sigma} \, d\Omega = \int_\Omega \vec{\omega} \cdot \vec{b} \, d\Omega + \int_\Gamma \vec{\omega} \cdot \vec{t} \, d\Gamma \tag{10.112}
$$

$$
\int_\Omega q \vec{\nabla} \cdot \vec{v} \, d\Omega = 0 \tag{10.113}
$$

where $\vec{\omega}$ and q are the velocity and pressure test functions respectively, and with

$$
\vec{\nabla}\vec{\omega} : \boldsymbol{\sigma} = \sum_{i,j}^{ndim} \frac{\partial \omega_i}{\partial x_j} \sigma_{ij}
$$

Assuming the Cauchy stress $\boldsymbol{\sigma}$ is given by the linear Stokes' law $\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau}$ with $\boldsymbol{\tau} = 2\eta \dot{\boldsymbol{\epsilon}}(\vec{v}) = \mathcal{C} : \vec{\nabla}\vec{v}$ where $\mathcal{C}$ is a fourth-order tensor with $\mathcal{C}_{ijkl} = \eta(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$. Then

$$
\begin{aligned}
\int_\Omega \vec{\nabla}\vec{\omega} : \boldsymbol{\sigma} \, d\Omega &= \int_\Omega \vec{\nabla}\vec{\omega} : (-p\mathbf{1} + \boldsymbol{\tau}) \, d\Omega \\
&= -\int_\Omega p \, \vec{\nabla}\vec{\omega} : \mathbf{1} \, d\Omega + \int_\Omega \vec{\nabla}\vec{\omega} : \boldsymbol{\tau} \, d\Omega \\
&= -\int_\Omega p \, \vec{\nabla} \cdot \vec{\omega} \, d\Omega + \int_\Omega \vec{\nabla}\vec{\omega} : \mathcal{C} : \vec{\nabla}\vec{v} \, d\Omega \tag{10.114}
\end{aligned}
$$

The weak form of the Stokes system now takes the form

$$-\int_\Omega p\, \vec{\nabla} \cdot \vec{\omega}\; d\Omega + \int_\Omega \vec{\nabla}\vec{\omega} : \mathcal{C} : \vec{\nabla}\vec{v}\; d\Omega \;=\; \int_\Omega \vec{\omega} \cdot \vec{b}\; d\Omega + \int_\Gamma \vec{\omega} \cdot \vec{t}\; d\Gamma \qquad (10.115)$$

$$\int_\Omega q\vec{\nabla} \cdot \vec{v}\; d\Omega \;=\; 0 \qquad (10.116)$$

Actually, the bilinear form with the two double dot products is not particularly convenient so it is always rewritten in terms of the strain rate vector

$$\vec{\dot{\varepsilon}}(\vec{v}) = \begin{pmatrix} \dot{\varepsilon}_{xx}(\vec{v}) \\ \dot{\varepsilon}_{yy}(\vec{v}) \\ \dot{\varepsilon}_{zz}(\vec{v}) \\ 2\dot{\varepsilon}_{xy}(\vec{v}) \\ 2\dot{\varepsilon}_{xz}(\vec{v}) \\ 2\dot{\varepsilon}_{yz}(\vec{v}) \end{pmatrix}$$

as one can easily show that

$$\vec{\nabla}\vec{\omega} : \mathcal{C} : \vec{\nabla}\vec{v} = \vec{\dot{\varepsilon}}(\vec{\omega})^T \cdot \boldsymbol{C}_\eta \cdot \vec{\dot{\varepsilon}}(\vec{v})$$

with

$$\boldsymbol{C}_\eta = \eta \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and since $\vec{\dot{\varepsilon}}(\vec{v}) = \boldsymbol{B} \cdot \vec{\mathcal{V}}$ and $\vec{\dot{\varepsilon}}(\vec{\omega}) = \boldsymbol{B} \cdot \vec{\mathcal{W}}$ then

$$\begin{aligned}
\int_\Omega \vec{\nabla}\vec{\omega} : \mathcal{C} : \vec{\nabla}\vec{v}\; d\Omega &= \int_\Omega \vec{\dot{\varepsilon}}(\vec{\omega})^T \cdot \boldsymbol{C}_\eta \cdot \vec{\dot{\varepsilon}}(\vec{v})\; d\Omega \\
&= \int_\Omega \vec{\mathcal{W}}^T \cdot \boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \cdot \vec{\mathcal{V}}\; d\Omega \\
&= \vec{\mathcal{W}}^T \cdot \int_\Omega \boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\; d\Omega \cdot \vec{\mathcal{V}} \\
&= \vec{\mathcal{W}}^T \cdot \mathbb{K} \cdot \vec{\mathcal{V}} \qquad (10.117)
\end{aligned}$$

Let us now turn to the mass conservation equation:

$$\int_\Omega q\vec{\nabla} \cdot \vec{v}\; d\Omega = \int_\Omega \vec{\mathcal{Q}}^T \vec{\mathcal{N}}^p \vec{\nabla} \cdot \vec{v}\; d\Omega = \vec{\mathcal{Q}}^T \cdot \int_\Omega \vec{\mathcal{N}}^p \vec{\nabla} \cdot \vec{v}\; d\Omega$$

We have $\vec{\nabla} \cdot \vec{v} = Tr[\dot{\varepsilon}(\vec{v})]$ but also

$$\begin{aligned}
\vec{\nabla} \cdot \vec{v} &= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \dot{\varepsilon}_{xx}(\vec{v}) \\ \dot{\varepsilon}_{yy}(\vec{v}) \\ \dot{\varepsilon}_{zz}(\vec{v}) \\ 2\dot{\varepsilon}_{xy}(\vec{v}) \\ 2\dot{\varepsilon}_{xz}(\vec{v}) \\ 2\dot{\varepsilon}_{yz}(\vec{v}) \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \vec{\dot{\varepsilon}}(\vec{v}) \\
&= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \boldsymbol{B} \cdot \vec{\mathcal{V}} \qquad (10.118)
\end{aligned}$$

so that

$$\vec{\mathcal{N}^p}\vec{\nabla}\cdot\vec{v} = \vec{\mathcal{N}^p} \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \boldsymbol{B} \cdot \vec{\mathcal{V}}$$
$$= \begin{pmatrix} \vec{\mathcal{N}^p} & \vec{\mathcal{N}^p} & \vec{\mathcal{N}^p} & 0 & 0 & 0 \end{pmatrix} \cdot \boldsymbol{B} \cdot \vec{\mathcal{V}}$$
$$= \boldsymbol{\mathcal{N}}^T \cdot \boldsymbol{B} \cdot \vec{\mathcal{V}} \tag{10.119}$$

Finally

$$\int_\Omega q \vec{\nabla} \cdot \vec{v} \, d\Omega = \vec{\mathcal{Q}}^T \cdot \int_\Omega \boldsymbol{\mathcal{N}}^T \cdot \boldsymbol{B} \, d\Omega \cdot \vec{\mathcal{V}}$$
$$= -\vec{\mathcal{Q}}^T \cdot \mathbb{G}^T \cdot \vec{\mathcal{V}} \tag{10.120}$$

where

$$\mathbb{G} = -\int_\Omega \boldsymbol{B}^T \cdot \boldsymbol{\mathcal{N}} \, d\Omega$$

Obviously the term $-\int_\Omega p \, \vec{\nabla}\vec{\omega} : \mathbf{1} \, d\Omega = -\int_\Omega p \, \vec{\nabla}\cdot\vec{\omega} \, d\Omega$ will take the form $\vec{\mathcal{P}}^T\cdot\boldsymbol{G}^T\cdot\vec{\mathcal{W}} = \vec{\mathcal{W}}^T\cdot\boldsymbol{G}\cdot\vec{\mathcal{P}}$ so that

$$-\int_\Omega p \, \vec{\nabla}\cdot\vec{\omega} \, d\Omega + \int_\Omega \vec{\nabla}\vec{\omega} : \mathcal{C} : \vec{\nabla}\vec{v} \, d\Omega = \vec{\mathcal{W}}^T \cdot \boldsymbol{G} \cdot \vec{\mathcal{P}} + \vec{\mathcal{W}}^T \cdot \mathbb{K} \cdot \vec{\mathcal{V}} = \vec{\mathcal{W}}^T \cdot \left(\boldsymbol{G} \cdot \vec{\mathcal{P}} + \mathbb{K} \cdot \vec{\mathcal{V}} \right)$$

The rhs are handled as shown previously. Since these relationships must work for any test function then it means that what multiplies $\vec{\mathcal{W}}$ and $\vec{\mathcal{Q}}$ must be null and we recover Eq. (10.96).

Note that this approach is quite versatile since it does not require to specify the coordinate system. The vector $\vec{\varepsilon}$ will contain the components of the strain rate tensor and in the end the matrix $\boldsymbol{B}$ will reflect the exact form of the strain rate tensor terms (see axisymmetric formulation).

10.8.4 Pressure scaling

As nicely explained in the step 32 of deal.ii[29], we often need to scale the $\mathbb{G}$ block since it is many orders of magnitude smaller than $\mathbb{K}$ (especially in geodynamics where viscosities are $\sim 10^{22}$), which introduces large inaccuracies in the solving process to the point that the solution is nonsensical. This scaling coefficient is η/L where η and L are representative viscosities and lengths. We start from

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix}$$

and introduce the scaling coefficient as follows (which in fact does not alter the solution at all):

$$\begin{pmatrix} \mathbb{K} & \frac{\eta}{L}\mathbb{G} \\ \frac{\eta}{L}\mathbb{G}^T & -\frac{\eta^2}{L^2}\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \frac{L}{\eta}\vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \frac{\eta}{L}\vec{h} \end{pmatrix}$$

We then end up with the modified Stokes system:

$$\begin{pmatrix} \mathbb{K} & \underline{\mathbb{G}} \\ \underline{\mathbb{G}}^T & \underline{\mathbb{C}} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \underline{\vec{\mathcal{P}}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \underline{\vec{h}} \end{pmatrix}$$

where

$$\underline{\mathbb{G}} = \frac{\eta}{L}\mathbb{G} \qquad \underline{\vec{\mathcal{P}}} = \frac{L}{\eta}\vec{\mathcal{P}} \qquad \underline{\mathbb{C}} = \frac{\eta^2}{L^2}\mathbb{C} \qquad \underline{\vec{h}} = \frac{\eta}{L}\vec{h}$$

After the solve phase, we recover the real pressure with $\vec{\mathcal{P}} = \frac{\eta}{L}\underline{\vec{\mathcal{P}}}$.

---

[29]https://www.dealii.org/9.0.0/doxygen/deal.II/step_32.html

10.8.5 Going from 3D to 2D

The world is three-dimensional. However, for many different reasons one may wish to solve problems which are two-dimensional.

Following ASPECT manual, we will think of two-dimensional models in the following way:

- We assume that the domain we want to solve on is a two-dimensional cross section (in the $x - y$ plane) that extends infinitely far in both negative and positive z direction.

- We assume that the velocity is zero in the z direction and that all variables have no variation in the z direction.

As a consequence, two-dimensional models are three-dimensional ones in which the z component of the velocity is zero and so are all z derivatives. This allows to reduce the momentum conservation equations from 3 equations to 2 equations. However, contrarily to what is often seen, the 3D definition of the deviatoric strain rate remains, i.e. in other words:

$$\dot{\varepsilon}^d = \dot{\varepsilon} - \frac{1}{3}(\vec{\nabla} \cdot \vec{v})\mathbf{1} \tag{10.121}$$

and not $1/2$. In light of all this, the full strain rate tensor and the deviatoric strain rate tensor in 2D are given by:

$$\varepsilon = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} & \dot{\varepsilon}_{yz} \\ \dot{\varepsilon}_{zx} & \dot{\varepsilon}_{zy} & \dot{\varepsilon}_{zz} \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & 0 \\ \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & \frac{\partial v}{\partial y} & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{10.122}$$

$$\dot{\varepsilon}^d = \frac{1}{3}\begin{pmatrix} 2\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} & \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & 0 \\ \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & -\frac{\partial u}{\partial x} + 2\frac{\partial v}{\partial y} & 0 \\ 0 & 0 & -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \end{pmatrix} \tag{10.123}$$

Although the bottom right term may be surprising, it is of no consequence when this expression of the deviatoric strain rate is used in the Stokes equation:

$$\vec{\nabla} \cdot 2\eta\dot{\varepsilon}^d =$$

FINISH!

In two dimensions the velocity is then $\vec{v} = (u, v)$ and the FEM building blocks and matrices are simply:

$$\vec{\dot{\varepsilon}} = \begin{pmatrix} \dot{\varepsilon}_{xx} \\ \dot{\varepsilon}_{yy} \\ 2\dot{\varepsilon}_{xy} \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial N_1^v}{\partial x} & 0 & \frac{\partial N_2^v}{\partial x} & 0 & \frac{\partial N_3^v}{\partial x} & 0 & \cdots & \frac{\partial N_{m_v}^v}{\partial x} & 0 \\ 0 & \frac{\partial N_1^v}{\partial y} & 0 & \frac{\partial N_2^v}{\partial y} & 0 & \frac{\partial N_3^v}{\partial y} & \cdots & 0 & \frac{\partial N_{m_v}^v}{\partial x} \\ \frac{\partial N_1^v}{\partial y} & \frac{\partial N_1^v}{\partial x} & \frac{\partial N_2^v}{\partial y} & \frac{\partial N_2^v}{\partial x} & \frac{\partial N_3^v}{\partial y} & \frac{\partial N_3^v}{\partial x} & \cdots & \frac{\partial N_{m_v}^v}{\partial y} & \frac{\partial N_{m_v}^v}{\partial x} \end{pmatrix}}_{\mathbf{B}} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ \cdots \\ u_{m_v} \\ v_{m_v} \end{pmatrix}}_{\vec{V}} \tag{10.124}$$

we have

$$\sigma_{xx} = -p + 2\eta\dot{\varepsilon}_{xx} \tag{10.125}$$
$$\sigma_{yy} = -p + 2\eta\dot{\varepsilon}_{yy} \tag{10.126}$$
$$\sigma_{xy} = +2\eta\dot{\varepsilon}_{xy} \tag{10.127}$$

so

$$\vec{\sigma} = -\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} p + \boldsymbol{C} \cdot \vec{\dot{\varepsilon}} = -\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \vec{N^p} \cdot \vec{P} + \boldsymbol{C} \cdot \boldsymbol{B} \cdot \vec{V} \tag{10.128}$$

with

$$\boldsymbol{C} = \eta \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{or} \qquad \boldsymbol{C} = \frac{\eta}{3} \begin{pmatrix} 4 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{10.129}$$

check the right C

Finally the matrix $\boldsymbol{N}^p$ is of size $3 \times m_p$:

$$\boldsymbol{N}^p = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \vec{N^p} = \begin{pmatrix} \vec{N^p} \\ \vec{N^p} \\ 0 \end{pmatrix} \tag{10.130}$$

10.8.6 The cylindrical axisymmetric case

In cylindrical coordinates the velocity gradient is given by

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & \frac{1}{r}\frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r} & \frac{\partial v_r}{\partial z} \\[2mm] \frac{\partial v_\theta}{\partial r} & \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r} & \frac{\partial v_\theta}{\partial z} \\[2mm] \frac{\partial v_z}{\partial r} & \frac{1}{r}\frac{\partial v_z}{\partial \theta} & \frac{\partial v_z}{\partial z} \end{pmatrix} \tag{10.131}$$

In the case of axisymmetry, and in this case symmetry about the z axis, there is invariance with respect to the rotation around the axis so stresses and other quantities are independent of the θ coordinate, or simply put $\partial_\theta \to 0$. The velocity gradient simplifies to:

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & -\frac{v_\theta}{r} & \frac{\partial v_r}{\partial z} \\[2mm] \frac{\partial v_\theta}{\partial r} & \frac{v_r}{r} & \frac{\partial v_\theta}{\partial z} \\[2mm] \frac{\partial v_z}{\partial r} & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix} \tag{10.132}$$

Also, it follows logically that $v_\theta = 0$ so that ultimately:

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & 0 & \frac{\partial v_r}{\partial z} \\[2mm] 0 & \frac{v_r}{r} & 0 \\[2mm] \frac{\partial v_z}{\partial r} & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix} \tag{10.133}$$

and the strain rate tensor is then given by

$$\dot{\varepsilon}(\vec{v}) = \frac{1}{2}\left(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T\right) = \begin{pmatrix} \frac{\partial v_r}{\partial r} & 0 & \frac{1}{2}\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right) \\[2mm] 0 & \frac{v_r}{r} & 0 \\[2mm] \frac{1}{2}\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right) & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix} \tag{10.134}$$

The velocity divergence $\vec{\nabla}\cdot\vec{v}$ is simply the trace of $\dot{\varepsilon}(\vec{v})$ so

$$\vec{\nabla}\cdot\vec{v} = \frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z}$$

The components of the $\vec{\dot{\varepsilon}}(\vec{v})$ vector are

$$\vec{\dot{\varepsilon}}(\vec{v}) = \begin{pmatrix} \dot{\varepsilon}_{rr} \\ \dot{\varepsilon}_{\theta\theta} \\ \dot{\varepsilon}_{zz} \\ 2\dot{\varepsilon}_{r\theta} \\ 2\dot{\varepsilon}_{rz} \\ 2\dot{\varepsilon}_{\theta z} \end{pmatrix} = \begin{pmatrix} \frac{\partial v_r}{\partial r} \\ \frac{v_r}{r} \\ \frac{\partial v_z}{\partial z} \\ 0 \\ \frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z} \\ 0 \end{pmatrix}$$

We see that there are two zeroes and consequently we only keep the four non zero components:

$$\vec{\dot{\varepsilon}}(\vec{v}) = \begin{pmatrix} \frac{\partial v_r}{\partial r} \\ \frac{v_r}{r} \\ \frac{\partial v_z}{\partial z} \\ \frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z} \end{pmatrix}$$

Only displacements in the r and z directions remain (note that $\dot{\varepsilon}_{\theta\theta}$ is in fact equal to v_r/r). In what follows I rename $u = v_r$ and $w = v_z$ to simplify notations. Then, inside an element we have

$$u^h(r,z) \; = \; \sum_{i=1}^{m_v} \mathcal{N}_i^v(r,z) u_i \tag{10.135}$$

$$w^h(r,z) \; = \; \sum_{i=1}^{m_v} \mathcal{N}_i^v(r,z) w_i \tag{10.136}$$

where $\mathcal{N}_i^v$ are the velocity basis functions attached to the m_v nodes of the element. We compute the elements of the $\vec{\dot{\varepsilon}}(\vec{v})$ vector as follows:

$$\dot{\varepsilon}_{rr} \; = \; \frac{\partial u^h}{\partial r} = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,z)\, u_i \tag{10.137}$$

$$\dot{\varepsilon}_{\theta\theta} \; = \; \frac{u_r^h}{r} = \frac{1}{r}\sum_{i=1}^{m} \mathcal{N}_i(r,z)\, u_i \tag{10.138}$$

$$\dot{\varepsilon}_{zz} \; = \; \frac{\partial w^h}{\partial z} = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial z}(r,z)\, w_i \tag{10.139}$$

$$2\dot{\varepsilon}_{rz} \; = \; \frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial r} = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial z}(r,z)u_i + \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,z)w_i \tag{10.140}$$

and then

$$\vec{\dot{\varepsilon}}^h = \begin{pmatrix} \frac{\partial u^h}{\partial r} \\[6pt] \frac{u^h}{r} \\[6pt] \frac{\partial w^h}{\partial z} \\[6pt] \frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial r} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial \mathcal{N}_1}{\partial r} & 0 & \frac{\partial \mathcal{N}_2}{\partial r} & 0 & \cdots & \cdots & \frac{\partial \mathcal{N}_{m_v}}{\partial r} & 0 \\[6pt] \frac{\mathcal{N}_1}{r} & 0 & \frac{\mathcal{N}_2}{r} & 0 & \cdots & \cdots & \frac{\mathcal{N}_{m_v}}{r} & 0 \\[6pt] 0 & \frac{\partial \mathcal{N}_1}{\partial z} & 0 & \frac{\partial \mathcal{N}_2}{\partial z} & \cdots & \cdots & 0 & \frac{\partial \mathcal{N}_{m_v}}{\partial z} \\[6pt] \frac{\partial \mathcal{N}_1}{\partial z} & \frac{\partial \mathcal{N}_1}{\partial r} & \frac{\partial \mathcal{N}_2}{\partial z} & \frac{\partial \mathcal{N}_2}{\partial r} & \cdots & \cdots & \frac{\partial \mathcal{N}_{m_v}}{\partial z} & \frac{\partial \mathcal{N}_{m_v}}{\partial r} \end{pmatrix}}_{\boldsymbol{B}(4 \times 2m_v)} \cdot \underbrace{\begin{pmatrix} u_1 \\ w_1 \\ u_2 \\ w_2 \\ \vdots \\ u_{m_v} \\ w_{m_v} \end{pmatrix}}_{\vec{\mathcal{V}}(2m_v \times 1)} \tag{10.141}$$

or $\vec{\dot{\varepsilon}}^h = \boldsymbol{B} \cdot \vec{\mathcal{V}}$, where $\vec{\mathcal{V}}$ is the vector of velocity dofs for an element.

Following the presentation of Section 10.8.3, we know that we will obtain

$$\underbrace{\left(-\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{\mathcal{N}}^p\, dV \right)}_{\mathbb{G}_e} \cdot \vec{P} + \underbrace{\left(\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{C} \cdot \boldsymbol{B}\, dV \right)}_{\mathbb{K}_e} \cdot \vec{V} = \underbrace{\int_{\Omega_e} \vec{\mathcal{N}}_b\, dV}_{\vec{f}_e} \tag{10.142}$$

with

$$\boldsymbol{C}_\eta = \eta \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

> **Note:** We have in cylindrical coordinates $dV = r dr d\theta dz$. The integral over the θ coordinate yields a factor 2π so for instance
>
> $$\mathbb{K}_e = 2\pi \iint_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{C} \cdot \boldsymbol{B} \; r dr dz \qquad (10.143)$$
>
> Note the r term in the integrand. The integration can now be performed as simply as was the case in the plane strain problem.

Note that it is common to actually start from $-\vec{\nabla} \cdot \vec{v} = 0$ (see Eq. (3) in [897]) so as to arrive at $\mathbb{G}_e^T \cdot \vec{V} = \vec{0}$. Ultimately we obtain the following system for each element:

$$\left(\begin{array}{cc} \mathbb{K}_e & \mathbb{G}_e \\ \mathbb{G}_e^T & 0 \end{array} \right) \cdot \left(\begin{array}{c} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{array} \right) = \left(\begin{array}{c} \vec{f}_e \\ 0 \end{array} \right)$$

Such a matrix is then generated for each element and then must me assembled into the global F.E. matrix.

Unfortunately there is not much teaching/practical material to be found in the literature with regards to axisymmetric flow. For example Donea and Huerta [363] do not even mention this problem. In Hughes [637] we find:

$$\boldsymbol{\sigma} = \left\{ \begin{array}{c} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \\ \sigma_{33} \end{array} \right\} = \left\{ \begin{array}{c} \sigma_{rr} \\ \sigma_{zz} \\ \sigma_{rz} \\ \sigma_{\theta\theta} \end{array} \right\} \qquad (2.12.3)$$

$$\boldsymbol{\epsilon} = \left\{ \begin{array}{c} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \\ \epsilon_{33} \end{array} \right\} = \left\{ \begin{array}{c} \epsilon_{rr} \\ \epsilon_{zz} \\ 2\epsilon_{rz} \\ \epsilon_{\theta\theta} \end{array} \right\} \qquad (2.12.4)$$

The ordering emanates from the following generalization of Table 2.7.1:

I	J	i	k	j	l
1		1		1	
3		1		2	
3		2		1	
2		2		2	
4		3		3	

The **D** array takes on the following form:

$$\boldsymbol{D} = [D_{IJ}] = \underbrace{\left[\begin{array}{cc} \boldsymbol{D}_{33} & \boldsymbol{D}_3 \\ \boldsymbol{D}_3^T & D_{44} \end{array} \right]}_{4 \times 4} \qquad (2.12.5)$$

$$\boldsymbol{D}_{33} = \left[\begin{array}{ccc} D_{11} & D_{12} & D_{13} \\ & D_{22} & D_{23} \\ \text{symmetric} & & D_{33} \end{array} \right] \qquad (2.12.6)$$

$$\boldsymbol{D}_3 = \left\{ \begin{array}{c} D_{14} \\ D_{24} \\ D_{34} \end{array} \right\} \qquad (2.12.7)$$

where $D_{IJ} = c_{ijkl}$, in which the indices are related by the table. The $\boldsymbol{B}_a$-matrix takes on the form

$$\boldsymbol{B}_a = \left[\begin{array}{cc} N_{a,1} & 0 \\ 0 & N_{a,2} \\ N_{a,2} & N_{a,1} \\ \hline \dfrac{N_a}{r} & 0 \end{array} \right] \qquad (2.12.8)$$

Again, a factor of $2\pi r$ needs to be included in all integrands.

Remark
The *plane strain* case may be obtained from the axisymmetric formulation by

i. Ignoring the $2\pi r$ factors; and

ii. Ignoring the fourth row of $\boldsymbol{B}_a$ and the fourth row and column of $\boldsymbol{D}$.

Also check page 469 of Gresho & Sani's book, and see their remark on axisymmetric case for the N-S equations on page 545.

📥 Relevant Literature:

- Masahisa Tabata. "Finite element analysis of axisymmetric flow problems". In: *Zeitschrift für angewandte Mathematik und Mechanik* 76 (1996), pp. 171–174

- Vitoriano Ruas. "Mixed finite element methods with discontinuous pressures for the axisymmetric Stokes problem". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics* 83.4 (2003), pp. 249–264. DOI: 10.1002/zamm.200310032

- Young-Ju Lee and Hengguang Li. "Axisymmetric Stokes equations in polygonal domains: regularity and finite element approximations". In: *Computers & Mathematics with Applications* 64.11 (2012), pp. 3500–3521. DOI: 10.1016/j.camwa.2012.08.014

- TE Price. "Numerically exact integration of a family of axisymmetric finite elements". In: *Communications in numerical methods in engineering* 19.4 (2003), pp. 253–261. DOI: 10.1002/cnm.583

10.9 Solving the elastic equations

NOW BEING REWORKED IN OVERLEAF

In what follows $\vec{v}$ now stands for the displacement vector, i.e. with units of length, not velocity. As before, the displacement inside an element is given by

$$\vec{v}^h(\vec{r}) = \sum_{i=1}^{m_v} N_i(\vec{r}) \, \vec{v}_i \tag{10.144}$$

where N_i are the polynomial basis functions for the displacement. Pressure does not appear in the equations so this is not a case of mixed FE as for the viscous Stokes flow.

Other notations are sometimes used for Eqs.(10.144):

$$u^h(\vec{r}) = \vec{N} \cdot \vec{u} \qquad v^h(\vec{r}) = \vec{N} \cdot \vec{v} \qquad w^h(\vec{r}) = \vec{N} \cdot \vec{w} \tag{10.145}$$

where $\vec{v} = (u, v, w)$ and $\vec{N}$ is the vector containing all basis functions evaluated at location $\vec{r}$:

$$\vec{N}^v = (N_1(\vec{r}), N_2(\vec{r}), N_3(\vec{r}), \dots N_{m_v}(\vec{r})) \tag{10.146}$$
$$\vec{N}^p = \left(N_1^p(\vec{r}), N_2^p(\vec{r}), N_3^p(\vec{r}), \dots N_{m_p}^p(\vec{r})\right) \tag{10.147}$$

and with

$$\vec{u} = (u_1, u_2, u_3, \dots u_{m_v}) \tag{10.148}$$
$$\vec{v} = (v_1, v_2, v_3, \dots v_{m_v}) \tag{10.149}$$
$$\vec{w} = (w_1, w_2, w_3, \dots w_{m_v}) \tag{10.150}$$
$$\tag{10.151}$$

In three dimensions We start from

$$\boldsymbol{\sigma} = \lambda (\vec{\nabla} \cdot \vec{v}) \mathbf{1} + 2\mu \boldsymbol{\varepsilon}$$

where μ is the shear modulus and λ the Lamé parameter.

$$
\begin{aligned}
\sigma_{xx} &= (\lambda + 2\mu)\varepsilon_{xx} + \lambda\varepsilon_{yy} + \lambda\varepsilon_{zz} \\
\sigma_{yy} &= \lambda\varepsilon_{xx} + (\lambda + 2\mu)\varepsilon_{yy} + \lambda\varepsilon_{zz} \\
\sigma_{zz} &= \lambda\varepsilon_{xx} + \lambda\varepsilon_{yy} + (\lambda + 2\mu)\varepsilon_{zz} \\
\sigma_{xy} &= 2\mu\varepsilon_{xy} \\
\sigma_{xz} &= 2\mu\varepsilon_{xz} \\
\sigma_{yz} &= 2\mu\varepsilon_{yz}
\end{aligned}
\tag{10.152}
$$

or,

$$
\vec{\sigma} =
\begin{pmatrix}
\sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz}
\end{pmatrix}
=
\begin{pmatrix}
\lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\
\lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\
\lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\
0 & 0 & 0 & \mu & 0 & 0 \\
0 & 0 & 0 & 0 & \mu & 0 \\
0 & 0 & 0 & 0 & 0 & \mu
\end{pmatrix}
\cdot
\begin{pmatrix}
\varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ 2\varepsilon_{xy} \\ 2\varepsilon_{xz} \\ 2\varepsilon_{yz}
\end{pmatrix}
= \vec{\varepsilon}
$$

The rest of the procedure is pretty straightforward since it follows the same ideas as for the mixed viscous case, except that we here build the $\mathbb{K}$ matrix only as follows:

$$\mathbb{K} = \int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{D} \cdot \boldsymbol{B} \, dV$$

In two dimensions   The above relationships simplify to

$$\sigma_{xx} = (\lambda + 2\mu)\varepsilon_{xx} + \lambda\varepsilon_{yy} \tag{10.153}$$

$$\sigma_{yy} = \lambda\varepsilon_{xx} + (\lambda + 2\mu)\dot{\varepsilon}_{yy} \tag{10.154}$$

$$\sigma_{xy} = 2\mu\dot{\varepsilon}_{xy} \tag{10.155}$$

so

$$\vec{\sigma} = \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} = \begin{pmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ 2\varepsilon_{xy} \end{pmatrix} = \vec{\varepsilon}$$

The axisymmetric case

We start from

$$\boldsymbol{\sigma} = \lambda\vec{\nabla}\cdot\vec{u}\,\mathbf{1} + 2\mu\boldsymbol{\varepsilon}(\vec{u}) \tag{10.156}$$

In cylindrical coordinates the velocity gradient is given by

$$\vec{\nabla}\vec{u} = \begin{pmatrix} \dfrac{\partial u_r}{\partial r} & \dfrac{1}{r}\dfrac{\partial u_r}{\partial \theta} - \dfrac{u_\theta}{r} & \dfrac{\partial u_r}{\partial z} \\[2ex] \dfrac{\partial u_\theta}{\partial r} & \dfrac{1}{r}\dfrac{\partial u_\theta}{\partial \theta} + \dfrac{u_r}{r} & \dfrac{\partial u_\theta}{\partial z} \\[2ex] \dfrac{\partial u_z}{\partial r} & \dfrac{1}{r}\dfrac{\partial u_z}{\partial \theta} & \dfrac{\partial u_z}{\partial z} \end{pmatrix}$$

In the case of axisymmetry, and in this case symmetry about the z axis, there is invariance with respect to the rotation around the axis so stresses and other quantities are independent of the θ coordinate, or simply put $\partial_\theta \to 0$. The velocity gradient simplifies to:

$$\vec{\nabla}\vec{u} = \begin{pmatrix} \dfrac{\partial u_r}{\partial r} & -\dfrac{u_\theta}{r} & \dfrac{\partial u_r}{\partial z} \\[2ex] \dfrac{\partial u_\theta}{\partial r} & \dfrac{u_r}{r} & \dfrac{\partial u_\theta}{\partial z} \\[2ex] \dfrac{\partial u_z}{\partial r} & 0 & \dfrac{\partial u_z}{\partial z} \end{pmatrix}$$

Also, it follows logically that $u_\theta = 0$ so that ultimately:

$$\vec{\nabla}\vec{u} = \begin{pmatrix} \dfrac{\partial u_r}{\partial r} & 0 & \dfrac{\partial u_r}{\partial z} \\[2ex] 0 & \dfrac{u_r}{r} & 0 \\[2ex] \dfrac{\partial u_z}{\partial r} & 0 & \dfrac{\partial u_z}{\partial z} \end{pmatrix}$$

and the strain tensor is then given by

$$\boldsymbol{\varepsilon}(\vec{u}) = \frac{1}{2}\left(\vec{\nabla}\vec{u} + \vec{\nabla}\vec{u}^T\right) = \begin{pmatrix} \dfrac{\partial u_r}{\partial r} & 0 & \dfrac{1}{2}\left(\dfrac{\partial u_z}{\partial r} + \dfrac{\partial u_r}{\partial z}\right) \\[2ex] 0 & \dfrac{u_r}{r} & 0 \\[2ex] \dfrac{1}{2}\left(\dfrac{\partial u_z}{\partial r} + \dfrac{\partial u_r}{\partial z}\right) & 0 & \dfrac{\partial u_z}{\partial z} \end{pmatrix} \tag{10.157}$$

The term $\vec{\nabla}\cdot\vec{u}$ is simply the trace of $\boldsymbol{\varepsilon}(\vec{u})$ so

$$\vec{\nabla}\cdot\vec{u} = \frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{\partial u_z}{\partial z}$$

Finally the full stress tensor is then

$$
\boldsymbol{\sigma} \;=\; \begin{pmatrix} \lambda(\frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{\partial u_z}{\partial z}) + 2\mu\frac{\partial u_r}{\partial r} & 0 & \mu(\frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}) \\ \\ 0 & \lambda(\frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{\partial u_z}{\partial z}) + 2\mu\frac{u_r}{r} & 0 \\ \\ \mu(\frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}) & 0 & \lambda(\frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{\partial u_z}{\partial z}) + 2\mu\frac{\partial u_z}{\partial z} \end{pmatrix}
$$

$$
\;=\; \begin{pmatrix} (\lambda+2\mu)\frac{\partial u_r}{\partial r} + \lambda(\frac{u_r}{r} + \frac{\partial u_z}{\partial z}) & 0 & \mu(\frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}) \\ \\ 0 & (\lambda+2\mu)\frac{u_r}{r} + \lambda(\frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z}) & 0 \\ \\ \mu(\frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}) & 0 & (\lambda+2\mu)\frac{\partial u_z}{\partial z} + \lambda(\frac{\partial u_r}{\partial r} + \frac{u_r}{r}) \end{pmatrix}
$$

As we did in the 2D case, we rewrite the six independent stress terms in to a vector $\vec{\sigma}$ and we use Eq. (10.156) to arrive at:

$$
\vec{\sigma} = \begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{r\theta} \\ \sigma_{rz} \\ \sigma_{\theta z} \end{pmatrix} = \begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{r\theta} \\ 2\varepsilon_{rz} \\ 2\varepsilon_{\theta z} \end{pmatrix} = \vec{\varepsilon}(\vec{u})
$$

or $\vec{\sigma} = \boldsymbol{D} \cdot \vec{\varepsilon}(\vec{u})$. Notice the similarity of matrix $\boldsymbol{D}$ with the one of Section (XXX) in the 3D penalty formulation case. The components of the $\vec{\varepsilon}$ vector are

$$
\vec{\varepsilon}(\vec{u}) = \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{r\theta} \\ 2\varepsilon_{rz} \\ 2\varepsilon_{\theta z} \end{pmatrix} = \begin{pmatrix} \frac{\partial u_r}{\partial r} \\ \frac{u_r}{r} \\ \frac{\partial u_z}{\partial z} \\ 0 \\ \frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z} \\ 0 \end{pmatrix}
$$

We see that there are two zeroes and consequently we'll find that $\sigma_{r\theta}$ and $\sigma_{\theta z}$ are also identically zero, so we discard these and end up with only four stress components :

$$
\vec{\sigma} = \begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix} = \begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{rz} \end{pmatrix}
$$

Note that in the literature the above relationship is often written

$$
\begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \lambda & \nu & 0 \\ \nu & 1-\nu & \nu & 0 \\ \nu & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & (1-2\nu)/2 \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{rz} \end{pmatrix}
$$

which is equivalent since $E = 2\mu(1+\nu)$ and $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}$ (see for instance Section 5.2.4 in [1540]).

Only displacements in the r and z directions remain (note that $\varepsilon_{\theta\theta}$ is in fact equal to u_r/r). In what follows I rename $u = u_r$ and $u_z = w$ to simplify notations. Then, inside an element we have

$$
\begin{aligned} u^h(r,z) &= \sum_{i=1}^{m} N_i(r,z) u_i \\ w^h(r,z) &= \sum_{i=1}^{m} N_i(r,z) w_i \end{aligned} \tag{10.158}
$$

where N_i are the basis functions attached to the m nodes of the element. We compute the elements of the $\boldsymbol{\varepsilon}$ tensor of Eq. (10.157) as follows:

$$\varepsilon_{rr} = \frac{\partial u^h}{\partial r} = \sum_{i=1}^{m} \frac{\partial N_i}{\partial r}(r, z)\, u_i \tag{10.159}$$

$$\varepsilon_{\theta\theta} = \frac{u_r^h}{r} = \frac{1}{r} \sum_{i=1}^{m} N_i(r, z)\, u_i \tag{10.160}$$

$$\varepsilon_{zz} = \frac{\partial w^h}{\partial z} = \sum_{i=1}^{m} \frac{\partial N_i}{\partial z}(r, z)\, w_i \tag{10.161}$$

$$\varepsilon_{rz} = \frac{1}{2}\frac{\partial u^h}{\partial z} + \frac{1}{2}\frac{\partial w^h}{\partial r} = \sum_{i=1}^{m} \frac{\partial N_i}{\partial z}(r, z)u_i + \sum_{i=1}^{m} \frac{\partial N_i}{\partial r}(r, z)w_i \tag{10.162}$$

Let us take $m = 3$, i.e. linear triangles, for simplicity. Then the strain vector $\vec{\varepsilon}^h$ is given by

$$\vec{\varepsilon}^h = \begin{pmatrix} \frac{\partial u^h}{\partial r} \\[2mm] \frac{u^h}{r} \\[2mm] \frac{\partial w^h}{\partial z} \\[2mm] \frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial r} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial N_1}{\partial r} & 0 & \frac{\partial N_2}{\partial r} & 0 & \frac{\partial N_3}{\partial r} & 0 \\[2mm] \frac{N_1}{r} & 0 & \frac{N_2}{r} & 0 & \frac{N_3}{r} & 0 \\[2mm] 0 & \frac{\partial N_1}{\partial z} & 0 & \frac{\partial N_2}{\partial z} & 0 & \frac{\partial N_3}{\partial z} \\[2mm] \frac{\partial N_1}{\partial z} & \frac{\partial N_1}{\partial r} & \frac{\partial N_2}{\partial z} & \frac{\partial N_2}{\partial r} & \frac{\partial N_3}{\partial z} & \frac{\partial N_3}{\partial r} \end{pmatrix}}_{\boldsymbol{B}(4\times 6)} \cdot \underbrace{\begin{pmatrix} u1 \\ w1 \\ u2 \\ w2 \\ u3 \\ w3 \end{pmatrix}}_{\vec{U}(6\times 1)}$$

or $\vec{\varepsilon}^h = \boldsymbol{B} \cdot \vec{U}$ and finally

$$\underbrace{\begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix}}_{\vec{\sigma}} = \underbrace{\begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix}}_{\boldsymbol{D}} \cdot \underbrace{\begin{pmatrix} \frac{\partial N_1}{\partial r} & 0 & \frac{\partial N_2}{\partial r} & 0 & \frac{\partial N_3}{\partial r} & 0 \\[2mm] \frac{N_1}{r} & 0 & \frac{N_2}{r} & 0 & \frac{N_3}{r} & 0 \\[2mm] 0 & \frac{\partial N_1}{\partial z} & 0 & \frac{\partial N_2}{\partial z} & 0 & \frac{\partial N_3}{\partial z} \\[2mm] \frac{\partial N_1}{\partial z} & \frac{\partial N_1}{\partial r} & \frac{\partial N_2}{\partial z} & \frac{\partial N_2}{\partial r} & \frac{\partial N_3}{\partial z} & \frac{\partial N_3}{\partial r} \end{pmatrix}}_{\boldsymbol{B}(4\times 6)} \cdot \underbrace{\begin{pmatrix} u1 \\ w1 \\ u2 \\ w2 \\ u3 \\ w3 \end{pmatrix}}_{\vec{U}(6\times 1)}$$

or,

$$\boxed{\vec{\sigma} = \boldsymbol{D} \cdot \boldsymbol{B} \cdot \vec{U}}$$

Note that in 2D, the matrix $\boldsymbol{D}$ is 3×3 and $\boldsymbol{B}$ is 3×6.

I do not know yet how to arrive at what follows

The 6×6 stiffness matrix is then

$$\mathbb{K} = \iiint \boldsymbol{B}^T \cdot \boldsymbol{D} \cdot \boldsymbol{B}\, dV$$

with $dV = r\,dr\,d\theta\,dz$ in cylindrical coordinates. The integral over the θ coordinate yields a factor 2π so

$$\mathbb{K} = 2\pi \iint \boldsymbol{B}^T \cdot \boldsymbol{D} \cdot \boldsymbol{B}\, r\,dr\,dz$$

The integration can now be performed as simply as was the case in the plane stress problem.

Note that in practice the matrix $\boldsymbol{D}$ is computed as follows (see for example Stone 63):

$$\boldsymbol{D} = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix} = \lambda \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \mu \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The divergence of the stress tensor is given by

$$\vec{\nabla} \cdot \boldsymbol{\sigma} = \left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{rr}) + \frac{1}{r}\frac{\partial\,\sigma_{r\theta}}{\partial\theta} + \frac{\partial\,\sigma_{rz}}{\partial z} - \frac{\sigma_{\theta\theta}}{r} \right] \vec{e}_r \tag{10.163}$$

$$+ \left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{r\theta}) + \frac{1}{r}\frac{\partial\,\sigma_{\theta\theta}}{\partial\theta} + \frac{\partial\,\sigma_{\theta z}}{\partial z} + \frac{\sigma_{r\theta}}{r} \right] \vec{e}_\theta \tag{10.164}$$

$$+ \left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{rz}) + \frac{1}{r}\frac{\partial\,\sigma_{\theta z}}{\partial\theta} + \frac{\partial\,\sigma_{zz}}{\partial z} \right] \vec{e}_z \tag{10.165}$$

Since $\sigma_{r\theta} = \sigma_{\theta r} = 0$ and $\sigma_{z\theta} = \sigma_{\theta z} = 0$ and since $\partial_\theta \to 0$ then

$$\vec{\nabla} \cdot \boldsymbol{\sigma} = \left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{rr}) + \frac{\partial\,\sigma_{rz}}{\partial z} - \frac{\sigma_{\theta\theta}}{r} \right] \vec{e}_r \tag{10.166}$$

$$+ \left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{rz}) + \frac{\partial\,\sigma_{zz}}{\partial z} \right] \vec{e}_z \tag{10.167}$$

Then

$$\vec{\nabla} \cdot \boldsymbol{\sigma}\big|_r = \frac{1}{r}\frac{\partial}{\partial r}(r\,\sigma_{rr}) + \frac{\partial\,\sigma_{rz}}{\partial z} - \frac{\sigma_{\theta\theta}}{r} \tag{10.168}$$

$$= \frac{\partial\sigma_{rr}}{\partial r} + \frac{1}{r}(\sigma_{rr} - \sigma_{\theta\theta}) + \frac{\partial\sigma_{rz}}{\partial z} \tag{10.169}$$

$$= \frac{\partial\sigma_{rr}}{\partial r} + \frac{2\mu}{r}\left(\frac{\partial\,u_r}{\partial r} - \frac{u_r}{r}\right) + \frac{\partial\sigma_{rz}}{\partial z} \tag{10.170}$$

$$\vec{\nabla} \cdot \boldsymbol{\sigma}\big|_z = \frac{\partial\sigma_{rz}}{\partial r} + \frac{\sigma_{rz}}{r} + \frac{\partial\,\sigma_{zz}}{\partial z} \tag{10.171}$$

10.10 The case against the $Q_1 \times P_0$ element

What follows was written by Dave May and sent to me by email in May 2014. It captures so well the problem at hand that I have decided to reproduce it hereunder.

In the case of the incompressible Stokes equations, we would like to solve

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ 0 \end{pmatrix}
$$

with an iterative method which is algorithmically scalable and optimal. Scalable here would mean that the number of iterations doesn't grow as the mesh is refined. Optimal means the solution time varies linearly with the total number of unknowns. When using a stable element, If we right precondition the above system with

$$
P = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ 0 & -\mathbb{S} \end{pmatrix}
$$

then convergence will occur in 2 iterations, however this requires an exact solve on $\mathbb{K}$ and on $\mathbb{S} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G}$ ($\mathbb{S}$ is the pressure schur complement). In practice, people relax the ideal "two iteration" scenario by first replacing $\mathbb{S}$ via $\mathbb{S}^* = \int \eta^{-1} \vec{N}^T \vec{N} \, dv$ (e.g. the pressure mass matrix scaled by the local inverse of viscosity).

$$
P^* = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ 0 & -\mathbb{S}^* \end{pmatrix}
$$

Using P^*, we obtain iteration counts which are larger than 2, but likely less than 10 - *however*, the number of iterations is independent of the mesh size. Replacing the exact $\mathbb{K}$ solve in P^* again increases the iterations required to solve Stokes, but it's still independent of the number of elements. When you have this behaviour, we say the preconditioner (P^*) is spectrally equivalent to the operator (which here is Stokes)

The problem with $Q_1 \times P_0$ is that there are no approximations for $\mathbb{S}$ which can be generated that ensure a spectrally equivalent P^*. Thus, as you refine the mesh using $Q_1 \times P_0$ elements, the iteration count ALWAYS grows. I worked on this problem during my thesis, making some improvements to the situation - however the problem still remains, it cannot be completely fixed and stems entirely from using unstable elements.

Citcom solvers works like this:

1. Solve $\mathbb{S} \cdot \mathcal{P} = \vec{f'}$ for pressure

2. Solve $\mathbb{K} \cdot \mathcal{V} = \vec{f} - \mathbb{G} \cdot \mathcal{P}$ for velocity

To obtain a scalable method, we need the number of iterations performed in (1) and (2) to be independent of the mesh. This means we need a spectrally equivalent preconditioner for $\mathbb{S}$ and $\mathbb{K}$. Thus, we have the same issue as when you iterate on the full stokes system.

When we don't have a scalable method, it means increasing the resolution requires more cpu time in a manner which cannot be predicted. The increase in iteration counts as the mesh is refined can be dramatic.

If we can bound the number of iterations, AND ensure that the cost per iteration is linearly related to the number of unknowns, then we have a good method which can run on any mesh resolution with a predictable cpu time. Obtaining scalable and optimal preconditioners for $\mathbb{K}$ is somewhat easier. Multi-grid will provide us with this.

The reason citcom doesn't run with 400^3 elements is exactly due to this issue. I've added petsc support in citcom (when i was young and naive) - but the root cause of the non-scalable solve is directly caused by the element choice. Note that many of the high resolution citcom jobs are single time step calculations— there is a reason for that.

For many lithosphere dynamics problems, we need a reasonable resolution (at least 200^3 and realistically 400^3 to 800^3). Given the increase in cost which occurs when using Q1P0, this is not achievable, as the citcom code has demonstrated. Note that citcom is 20 years old now and for its time, it was great, but we know much more now and we know how to improve on it. As a result of this realization, I dumped all my old Q1P0 codes (and Q1Q1 codes, but for other reasons) in the trash and started from scratch. The only way to make something like 800^3 tractable is via iterative, scalable and optimal methods and that mandates stable elements. I can actually run at something like 1000^3 (nodal points) these days because of such design choices.

10.11 Isoviscous Stokes for incompressible flow

We start from the momentum equation:

$$-\vec{\nabla}p + \vec{\nabla}\cdot(2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v})) + \rho\vec{g} = \vec{0} \tag{10.172}$$

When the viscosity is constant in space, it can be taken out of the divergence operator:

$$-\vec{\nabla}p + 2\eta\vec{\nabla}\cdot\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) + \rho\vec{g} = \vec{0} \tag{10.173}$$

Let us for simplicity look at a 2D Cartesian formulation of this equation and for incompressible flow:

$$
\begin{aligned}
2\vec{\nabla}\cdot\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) &= \vec{\nabla}\cdot\left(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T\right) \\
&= (\partial_x\ \partial_y)\cdot\begin{pmatrix}\partial_x u & \partial_x v \\ \partial_y u & \partial_y v\end{pmatrix} + (\partial_x\ \partial_y)\cdot\begin{pmatrix}\partial_x u & \partial_y u \\ \partial_x v & \partial_y v\end{pmatrix} \\
&= (\partial_x^2 u + \partial_y^2 u, \partial_x^2 v + \partial_y^2 v) + (\partial_x\partial_x u + \partial_y\partial_x v, \partial_x\partial_y u + \partial_y\partial_y v) \\
&= (\partial_x^2 u + \partial_y^2 u, \partial_x^2 v + \partial_y^2 v) + (\partial_x\underbrace{(\partial_x u + \partial_y v)}_{=0}, \partial_y\underbrace{(\partial_x u + \partial_y v)}_{=0}) \\
&= (\partial_x^2 u + \partial_y^2 u, \partial_x^2 v + \partial_y^2 v)
\end{aligned} \tag{10.174}
$$

and then finally the Stokes equation is:

$$-\vec{\nabla}p + \eta\Delta\vec{v} + \rho\vec{g} = \vec{0} \tag{10.175}$$

The mass conservation equation remains unchanged and so does the pressure gradient term. We shall then focus on the weak form of the previously obtained term. We multiply it by a velocity test function $\mathcal{N}_i^v$ and integrate over an element[30]:

$$
\begin{aligned}
&\int_{\Omega_e}\mathcal{N}_i^v\Delta\vec{v}^h dV \\
&= \int_{\Omega_e}\begin{pmatrix}\mathcal{N}_i^v\Delta u^h \\ \mathcal{N}_i^v\Delta v^h\end{pmatrix}dV \\
&= \int_{\Omega_e}\begin{pmatrix}\mathcal{N}_i^v\vec{\nabla}\cdot\vec{\nabla}u^h \\ \mathcal{N}_i^v\vec{\nabla}\cdot\vec{\nabla}v^h\end{pmatrix}dV \\
&= \int_{\Omega_e}\begin{pmatrix}\vec{\nabla}\mathcal{N}_i^v\cdot\vec{\nabla}u^h \\ \vec{\nabla}\mathcal{N}_i^v\cdot\vec{\nabla}v^h\end{pmatrix}dV \\
&= \int_{\Omega_e}\begin{pmatrix}\partial_x\mathcal{N}_i^v\partial_x u^h + \partial_y\mathcal{N}_i^v\partial_y u^h \\ \partial_x\mathcal{N}_i^v\partial_x v^h + \partial_y\mathcal{N}_i^v\partial_y v^h\end{pmatrix}dV \\
&= \int_{\Omega_e}\begin{pmatrix}\partial_x\mathcal{N}_i^v & \partial_y\mathcal{N}_i^v & 0 & 0 \\ 0 & 0 & \partial_x\mathcal{N}_i^v & \partial_y\mathcal{N}_i^v\end{pmatrix}\cdot\begin{pmatrix}\partial_x u^h \\ \partial_y u^h \\ \partial_x v^h \\ \partial_y v^h\end{pmatrix}dV \\
&= \int_{\Omega_e}\begin{pmatrix}\partial_x\mathcal{N}_i^v & \partial_y\mathcal{N}_i^v & 0 & 0 \\ 0 & 0 & \partial_x\mathcal{N}_i^v & \partial_y\mathcal{N}_i^v\end{pmatrix}\cdot\begin{pmatrix}\partial_x\mathcal{N}_1^v & 0 & \partial_x\mathcal{N}_2^v & 0 & \cdots & \partial_x\mathcal{N}_{m_v}^v & 0 \\ \partial_y\mathcal{N}_1^v & 0 & \partial_y\mathcal{N}_2^v & 0 & \cdots & \partial_y\mathcal{N}_{m_v}^v & 0 \\ 0 & \partial_x\mathcal{N}_1^v & 0 & \partial_x\mathcal{N}_2^v & \cdots & 0 & \partial_x\mathcal{N}_{m_v}^v \\ 0 & \partial_y\mathcal{N}_1^v & 0 & \partial_y\mathcal{N}_2^v & \cdots & 0 & \partial_y\mathcal{N}_{m_v}^v\end{pmatrix}\cdot\begin{pmatrix}u_1 \\ v_1 \\ u_2 \\ v_2 \\ \cdots \\ u_{m_v} \\ v_{m_v}\end{pmatrix}
\end{aligned}
$$

---

[30]As per usual we discard the surface term when integrating by parts

Writing this equation for $i = 1, ... m_v$, we obtain:

$$
\int
\begin{pmatrix}
\partial_x \mathcal{N}_1^\gamma & \partial_y \mathcal{N}_1^\gamma & 0 & 0 \\
0 & 0 & \partial_x \mathcal{N}_1^\gamma & \partial_y \mathcal{N}_1^\gamma \\
\partial_x \mathcal{N}_2^\gamma & \partial_y \mathcal{N}_2^\gamma & 0 & 0 \\
0 & 0 & \partial_x \mathcal{N}_2^\gamma & \partial_y \mathcal{N}_2^\gamma \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\partial_x \mathcal{N}_{m_v}^\gamma & \partial_y \mathcal{N}_{m_v}^\gamma & 0 & 0 \\
0 & 0 & \partial_x \mathcal{N}_{m_v}^\gamma & \partial_y \mathcal{N}_{m_v}^\gamma
\end{pmatrix}
\cdot
\begin{pmatrix}
\partial_x \mathcal{N}_1^\gamma & 0 & \partial_x \mathcal{N}_2^\gamma & 0 & \cdots & \partial_x \mathcal{N}_{m_v}^\gamma & 0 \\
\partial_y \mathcal{N}_1^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma & 0 & \cdots & \partial_y \mathcal{N}_{m_v}^\gamma & 0 \\
0 & \partial_x \mathcal{N}_1^\gamma & 0 & \partial_x \mathcal{N}_2^\gamma & \cdots & 0 & \partial_x \mathcal{N}_{m_v}^\gamma \\
0 & \partial_y \mathcal{N}_1^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma & \cdots & 0 & \partial_y \mathcal{N}_{m_v}^\gamma
\end{pmatrix}
\cdot
\underbrace{
\begin{pmatrix}
u_1 \\ v_1 \\ u_2 \\ v_2 \\ \cdots \\ u_{m_v} \\ v_{m_v}
\end{pmatrix}
}_{\vec{V}}
$$

or,

$$
\boldsymbol{K}_\eta = \eta \int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{B} \; dV
$$

where $\boldsymbol{B}$ is a $(ndim * ndim) \times (m_v * ndofV)$ matrix (see also Eq. 6.24 of Donea and Huerta [363]).

In three dimensions, the matrix $\boldsymbol{B}$ is given by

$$
\begin{pmatrix}
\partial_x \mathcal{N}_1^\gamma & 0 & \partial_x \mathcal{N}_2^\gamma & 0 & \cdots & \partial_x \mathcal{N}_{m_v}^\gamma & 0 \\
\partial_y \mathcal{N}_1^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma & 0 & \cdots & \partial_y \mathcal{N}_{m_v}^\gamma & 0 \\
\partial_z \mathcal{N}_1^\gamma & 0 & \partial_z \mathcal{N}_2^\gamma & 0 & \cdots & \partial_z \mathcal{N}_{m_v}^\gamma & 0 \\
0 & \partial_x \mathcal{N}_1^\gamma & 0 & \partial_x \mathcal{N}_2^\gamma & \cdots & 0 & \partial_x \mathcal{N}_{m_v}^\gamma \\
0 & \partial_y \mathcal{N}_1^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma & \cdots & 0 & \partial_y \mathcal{N}_{m_v}^\gamma \\
0 & \partial_z \mathcal{N}_1^\gamma & 0 & \partial_z \mathcal{N}_2^\gamma & \cdots & 0 & \partial_z \mathcal{N}_{m_v}^\gamma
\end{pmatrix}
$$

10.12 $Q_1 \times P_0$ macro-elements

The Stenberg macro-element

This macro-element is introduced in Stenberg [1288] (1984).

$\bullet\ \vec{v}\quad \bigcirc p$
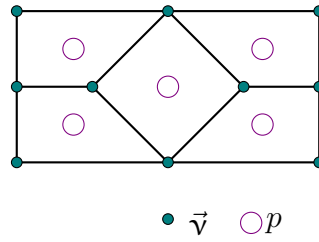
Gresho & Sani [517] state: "For fans of $Q_1 \times Q_0$ who want guaranteed optimal convergence of both u and p (with however larger error constants caused by the distorted shapes?), one way to assure this is to discretise via the macro elements above, each composed of five $Q_1 \times Q_0$ quadrilaterals. Such checkerboard-killer meshes have been employed in practice by (at least) Bathé [237]. Both the macro-element and the proof are due to Stenberg [1288]."

Chapelle & Bathe [237]: "the numerical inf-sup test is passed for this mesh and in fact, this behavior was proven analytically (see Brezzi & Fortin [157], see also Le Tallec & Ruas [820]).
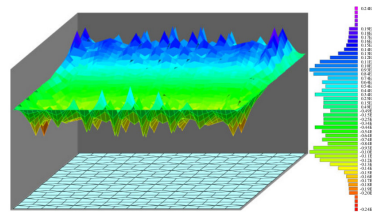
Figure 14. Error of component 1 of u_h on the Stenberg mesh [15].

Taken from Qin & Zhang (2007) [1091].

Implemented in STONE 78.

📚 Relevant Literature: Fig 3.12 of Elman *et al.* book [394]. Mentioned in Qin and Zhang [1092] (2007).

The Le Tallec macro-element

This macro-element is introduced in LeTallec [819] (1981).

$\bullet\ \vec{v}\quad \bigcirc p$
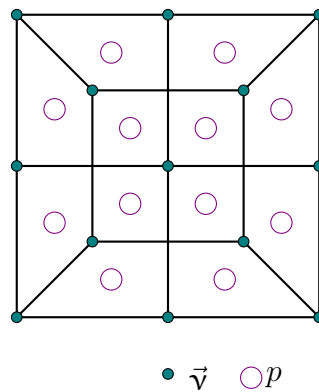
This macro-element has been proven stable in [819, 820], i.e. it satisfies the stability condition (see Section **??**). It is also mentioned in Qin & Zhang (2007) [1091].

Implemented in STONE 78.

The Qin & Zhang macro-elements

In their paper Qin and Zhang [1091] (2007) the authors mention the Stenberg and Le Tallec macro-elements and also introduce three new ones:

$\bullet\ \vec{v}\quad\bigcirc p$ $\bullet\ \vec{v}\quad\bigcirc p$ $\bullet\ \vec{v}\quad\bigcirc p$
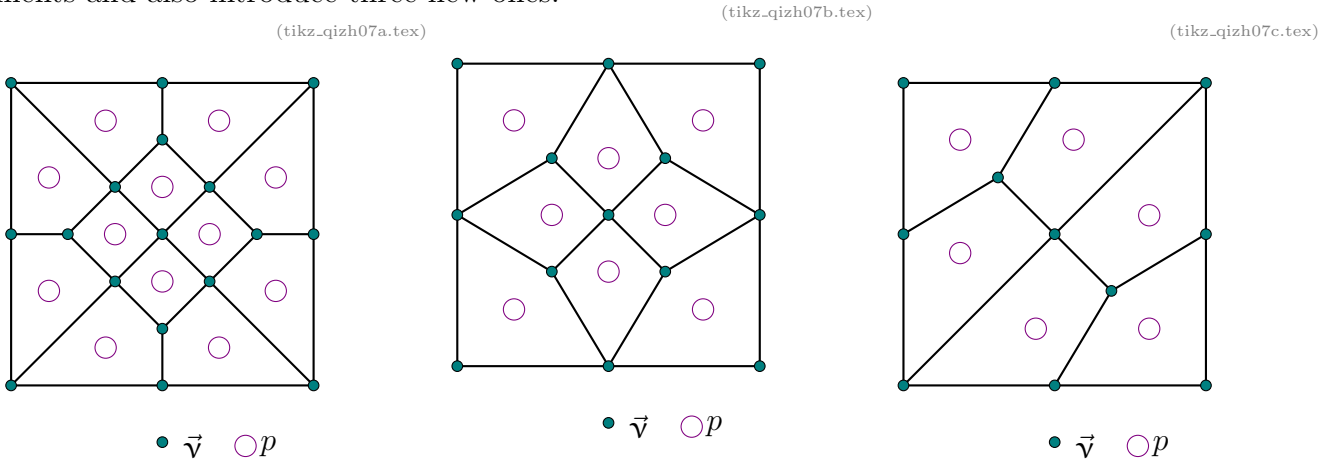
They also indicate that although stable, these macro-elements are inferior to the above two (Stenberg & Le Tallec).

New macro-elements ?
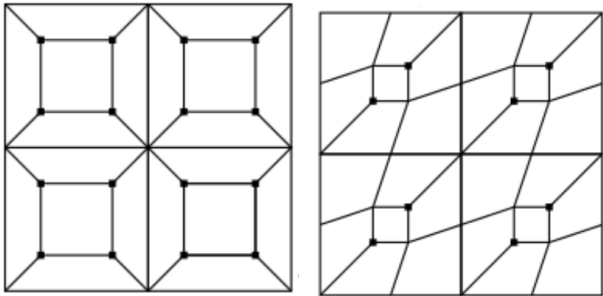
I came up with these, no idea whether these are stable/usable or better than the others.

490

10.13 Solving the Stokes system

Let us start again from the (full) Stokes system:

$$\underbrace{\left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{array} \right)}_{\mathcal{A}} \cdot \left(\begin{array}{c} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{array} \right) = \left(\begin{array}{c} \vec{f} \\ \vec{h} \end{array} \right) \tag{10.176}$$

We need to solve this system in order to obtain the solution, i.e. the $\vec{\mathcal{V}}$ and $\vec{\mathcal{P}}$ vectors. But how? Unfortunately, this question is not simple to answer and the appropriate method depends on many parameters, but mainly on how big the matrix blocks are and what the condition number of the matrix $\mathbb{K}$ is.

First let us start with an obvious question: couldn't we just compute the inverse of the matrix $\mathcal{A}$? Under the assumption that the inverse of $\mathbb{K}$ and $\mathbb{S}$ exists, we can and we find[31]

$$\mathcal{A}^{-1} = \left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{array} \right)^{-1} = \left(\begin{array}{cc} \mathbb{K}^{-1} + \mathbb{K}^{-1} \cdot \mathbb{G} \cdot \mathbb{S}^{-1} \cdot \mathbb{G}^T \cdot \mathbb{K}^{-1} & -\mathbb{K}^{-1} \cdot \mathbb{G} \cdot \mathbb{S}^{-1} \\ -\mathbb{S}^{-1} \cdot \mathbb{G}^T \cdot \mathbb{K}^{-1} & \mathbb{S}^{-1} \end{array} \right)$$

However, such an expression is of limited interest in the numerical solution of saddle point problems since it showcases 5 times the inverse of $\mathbb{K}$ and more importantly the inverse of the Schur complement matrix § which is likely to be a full matrix so that we never want to compute it explicitely.

As concisely explained in Clevenger & Heister (2021) [279], there are three common approaches used in the literature for solving the above equation on large scales:

- a pressure corrected, Schur complement CG scheme, using multigrid as an approximation to the velocity block;

- a block-preconditioned Krylov method, also using multigrid on the velocity block. For this method, there are two main types:

  - GMRES[894, 1161] (or any Krylov method not requiring symmetry) with block-triangular preconditioner (This is what ASPECT does):

$$\boldsymbol{P} = \left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ 0 & -\mathbb{S} \end{array} \right)$$

  - MINRES[498] with block-diagonal preconditioner

$$\boldsymbol{P} = \left(\begin{array}{cc} \mathbb{K} & 0 \\ 0 & -\mathbb{S} \end{array} \right)$$

- an all-at-once multigrid performed on the entire Stokes system, using Uzawa-type smoothers.

🔖 Relevant Literature: Preconditioners for Incompressible Navier-Stokes Solvers [1225]
Saddle point preconditioners have been extensively discussed and studied [78], [998]
Diagonal preconditioners in [1222], [52].
Pragmatic solvers for 3D Stokes problems with heterogeneous coefficients [1182]

---

[31]The matrix $\mathbb{C}$ is here omitted but it bears no consequences on the conclusion.

10.13.1 When using the penalty formulation

In this case we are only solving for velocity since pressure has been eliminated and is later recovered in a post-processing step:

$$(\mathbb{K}_\eta + \mathbb{K}_\lambda) \cdot \vec{\mathcal{V}} = \vec{f}$$

We also know that the penalty factor λ is many orders of magnitude higher than the viscosity and in combination with the use of the $Q_1 \times P_0$ element the resulting matrix condition number is very high so that the use of iterative solvers is precluded. Indeed codes such as SOPALE [452], DOUAR [144], FANTOM [1343] or SULEC [1094] relying on the penalty formulation all use direct solvers. The most popular are BLKFCT[32], MUMPS[33][13, 15, 14, 16, 12], PaSTiX [596], WSMP[34] [540, 541], UMFPACK and CHOLMOD[35], SuperLU[36], PARDISO[37] [340, 1417, 769], or those inside PETSc[38].

Braun *et al.* (2008) [144] list the following features of direct solvers:

- Robust

- Black-box operation

- Difficult to parallelize

- Memory consumption

- Limited scalability

The main advantage of direct solvers is used in this case: They can solve ill-conditioned matrices. However, memory requirements for the storage of number of nonzeros in the Cholesky matrix grow very fast as the number of equations/grid size increases, especially in 3D, to the point that even modern computers with tens of Gb of RAM cannot deal with a $\sim 100^3$ element mesh. This explains why direct solvers are often used for 2D problems and rarely in 3D with noticeable exceptions [1347, 1478, 145, 853, 10, 9, 11, 1452, 990].

Note that Pedroso [1044] (2024) conducted a detailed study comapring MUMPS, UMFPACK, and Intel DSS (PARDISO). The conclusions are

1. Intel DSS is not ready for production codes because it computes incorrect results without warnings;

2. UMFPACK reaches the memory limit earlier than the other solvers and must be configured with the automatic symmetry strategy to yield correct results; and

3. MUMPS is not thread-safe and, in particular, does not work well with OpenBLAS, which causes thread 'locks' and conflicts. However, MUMPS works well with Intel MKL and is the only solver able to tackle massive systems.

In Saramito [1188] we find the following table whigh gives the aymptotic computing time versus the size of the sparse matrix n and the geometry dimension d:

d	factorize	solve
1	n	n
2	$n^{3/2}$	$n \log n$
3	n^2	$n^{4/3}$

---

[32] http://dm.unife.it/blkfclt/

[33] http://mumps.enseeiht.fr/

[34] http://www.research.ibm.com/projects/wsmp

[35] http://faculty.cse.tamu.edu/davis/suitesparse.html

[36] https://portal.nersc.gov/project/sparse/superlu/

[37] https://www.pardiso-project.org/

[38] https://www.mcs.anl.gov/petsc/

10.13.2 Uzawa algorithms and the Schur complement approach

Let us write the above system as two equations:

$$\mathbb{K} \cdot \vec{\mathcal{V}} + \mathbb{G} \cdot \vec{\mathcal{P}} = \vec{f} \tag{10.177}$$

$$\mathbb{G}^T \cdot \vec{\mathcal{V}} - \mathbb{C} \cdot \vec{\mathcal{P}} = \vec{h} \tag{10.178}$$

The first line can be re-written $\vec{\mathcal{V}} = \mathbb{K}^{-1} \cdot (\vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}})$ and can be inserted in the second:

$$\mathbb{G}^T \cdot \vec{\mathcal{V}} = \mathbb{G}^T \cdot [\mathbb{K}^{-1} \cdot (\vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}})] - \mathbb{C} \cdot \vec{\mathcal{P}} = \vec{h} \tag{10.179}$$

or,

$$(\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot \vec{\mathcal{P}} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \vec{f} - \vec{h} \tag{10.180}$$

The matrix $\mathbb{S} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}$ is called the Schur complement. It is Symmetric (since $\mathbb{K}$ is symmetric) and Positive-Definite[39] (SPD) if $Ker(\mathbb{G}) = 0$. Having solved this equation (i.e. we have obtained $\vec{\mathcal{P}}$), the velocity can be recovered by solving $\mathbb{K} \cdot \vec{\mathcal{V}} = \vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}$.

Remark. *The Schur complement matrix naturally occurs when the Stokes matrix is decomposed using a LDU block-factorisation. Indeed, we have*

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & 0 \\ \mathbb{G}^T \cdot \mathbb{K}^{-1} & \boldsymbol{I} \end{pmatrix} \cdot \begin{pmatrix} \mathbb{K} & 0 \\ 0 & -\mathbb{S} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & \mathbb{K}^{-1} \cdot \mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix}$$

For now, let us assume that we have built the $\mathbb{S}$ matrix[40] and the right hand side $\underline{\vec{f}} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \vec{f} - \vec{h}$. We must solve $\mathbb{S} \cdot \vec{\mathcal{P}} = \underline{\vec{f}}$. It is easy to see that $\mathbb{S}$ is actually a full matrix (i.e. not sparse) and aside from the costs of building it explicitly using a direct solver would require a lot of memory so that we must then turn to iterative methods.

One can resort to so-called Richardson iterations, defined as follows (e.g., see Varga [1411], p141): in solving the matrix equation $\boldsymbol{A} \cdot \vec{X} = \vec{b}$, the Richardson iterative method is defined by:

$$\vec{X}_{k+1} = \vec{X}_k + \alpha_k (-\boldsymbol{A} \cdot \vec{X}_k + \vec{b}) \qquad m \geq 0 \tag{10.181}$$

where the α_k's are real scalars. It is easy to see that when the method converges then $\vec{X}_{k+1} \simeq \vec{X}_k$ and then for $\alpha_k \neq 0$ then $\boldsymbol{A} \cdot \vec{X} = \vec{b}$ is satisfied. In our case, it writes:

$$\begin{aligned} \vec{\mathcal{P}}_{k+1} &= \vec{\mathcal{P}}_k + \alpha_k (-\mathbb{S} \cdot \vec{\mathcal{P}}_k + \underline{\vec{f}}) \\ &= \vec{\mathcal{P}}_k + \alpha_k \left[-(\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot \vec{\mathcal{P}}_k + (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \vec{f} - \vec{h}) \right] \\ &= \vec{\mathcal{P}}_k + \alpha_k \left[\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot (-\mathbb{G} \cdot \vec{\mathcal{P}}_k + \vec{f}) - \mathbb{C} \cdot \vec{\mathcal{P}}_k - \vec{h} \right] \\ &= \vec{\mathcal{P}}_k + \alpha_k \left[\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot (\mathbb{K} \cdot \vec{\mathcal{V}}_k) - \mathbb{C} \cdot \vec{\mathcal{P}}_k - \vec{h} \right] \\ &= \vec{\mathcal{P}}_k + \alpha_k \left(\mathbb{G}^T \cdot \vec{\mathcal{V}}_k - \mathbb{C} \cdot \vec{\mathcal{P}}_k - \vec{h} \right) \end{aligned} \tag{10.182}$$

The above iterations are then carried out and for each new pressure field the associated velocity field is computed. The method of using Richardson iterations applied to the Schur complement is commonly called the Uzawa algorithm (see Braess [136, p221] [41]).

---

[39]M positive definite $\iff x^T M x > 0 \; \forall \, x \in \mathbb{R}^n \setminus \boldsymbol{0}$

[40]We will revisit this topic later on, but be aware that we never build $\mathbb{S}$ in practice.

[41]I have slightly altered the indices of the velocities wrt the book

This same algorithm is to be found on page 59 or [1188].

This method is rather simple to implement, although what makes an appropriate set of α_k values is not straightforward, which is why the conjugate gradient is often preferred, as detailed in the next section.

It is known that such iterations will converge for $0 < \alpha < \rho(\mathbb{S}) = \lambda_{max}(\mathbb{S})$ where $\rho(\mathbb{S})$ is the spectral radius of the matrix $\mathbb{S}$ which is essentially the largest, in absolute value, eigenvalue of $\mathbb{S}$ (neither of which can be computed easily). It can also be proven that the rate of convergence depends on the condition number of the matrix.

Richardson iterations are part of the family of stationary iterative methods[42], since it can be rewritten

$$\vec{X}_{k+1} = (\boldsymbol{I} - \alpha_k \boldsymbol{A}) \cdot \vec{X}_k + \alpha_k \vec{b} \tag{10.185}$$

which is the definition of a stationary method. The four main stationary methods are the Jacobi method, Gauss-Seidel method, successive overrelaxation method (SOR), and symmetric successive overrelaxation method (SSOR)

Since the α parameter is the key to a successful Uzawa algorithm, this issue has of course been looked into. What follows is presented in p221 of Braess [136]. For the analysis of the Uzawa algorithm, we define the residue

$$\vec{\mathcal{R}}_k = \vec{h} - \mathbb{G}^T \cdot \vec{\mathcal{V}}_k + \mathbb{C} \cdot \vec{\mathcal{P}}_k$$

In addition, suppose the solution of the saddle point problem is denoted by $(\vec{\mathcal{V}}^\star, \vec{\mathcal{P}}^\star)$ so that we have

$$\vec{f} = \mathbb{K} \cdot \vec{\mathcal{V}}^\star + \mathbb{G} \cdot \vec{\mathcal{P}}^\star \qquad \text{and} \qquad \vec{h} = \mathbb{G}^T \cdot \vec{\mathcal{V}}^\star - \mathbb{C} \cdot \vec{\mathcal{P}}^\star$$

Now substituting the iteration formula for $\mathcal{V}_k$, and inserting $\vec{f}$ and $\vec{h}$ from above, we get

$$\begin{aligned}
\vec{\mathcal{R}}_k &= \vec{h} - \mathbb{G}^T \cdot \vec{\mathcal{V}}_k + \mathbb{C} \cdot \vec{\mathcal{P}}_k \\
&= \vec{h} - \mathbb{G}^T \cdot \mathbb{K}^{-1}(\vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_k) + \mathbb{C} \cdot \vec{\mathcal{P}}_k \tag{10.186} \\
&= (\mathbb{G}^T \cdot \vec{\mathcal{V}}^\star - \mathbb{C} \cdot \vec{\mathcal{P}}^\star) - \mathbb{G}^T \cdot \mathbb{K}^{-1}(\mathbb{K} \cdot \vec{\mathcal{V}}^\star + \mathbb{G} \cdot \vec{\mathcal{P}}^\star - \mathbb{G} \cdot \vec{\mathcal{P}}_k) + \mathbb{C} \cdot \vec{\mathcal{P}}_k \tag{10.187} \\
&= (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot (\vec{\mathcal{P}}_k - \vec{\mathcal{P}}^\star) \tag{10.188}
\end{aligned}$$

From Eq. (10.184) it follows that:

$$\begin{aligned}
\vec{\mathcal{P}}_{k+1} - \vec{\mathcal{P}}_k &= \alpha \, (\mathbb{G}^T \cdot \vec{\mathcal{V}}_k - \mathbb{C} \cdot \vec{\mathcal{P}}_k - \vec{h}) \tag{10.189} \\
&= -\alpha \, \vec{\mathcal{R}}_k \tag{10.190} \\
&= -\alpha \, (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot (\vec{\mathcal{P}}_k - \vec{\mathcal{P}}^\star) \tag{10.191} \\
&= \alpha \, (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot (\vec{\mathcal{P}}^\star - \vec{\mathcal{P}}_k) \tag{10.192}
\end{aligned}$$

Thus the Uzawa algorithm is equivalent to applying the gradient method to the reduced equation using a fixed step size. In particular, the iteration converges for $\alpha < 2||\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}||^{-1}$ and one can show that the good step size α_k is given by

$$\alpha_k = \frac{\vec{\mathcal{R}}_k \cdot \vec{\mathcal{R}}_k}{(\mathbb{G} \cdot \vec{\mathcal{R}}_k) \cdot (\mathbb{K}^{-1} \cdot \mathbb{G} \cdot \vec{\mathcal{R}}_k)} \tag{10.193}$$

---

[42]https://mathworld.wolfram.com/StationaryIterativeMethod.html

However, if we were to use this rule formally, we would need an additional multiplication by $\mathbb{K}^{-1}$ in every step of the iteration. This can be avoided by storing an auxiliary vector. Note that this algorithm is presented in Zienkiewicz *et al.* (1985) [1544] in the context of viscoplastic flow.

As mentioned above, there is a way to rework the original Uzawa algorithm to include Eq. (10.193). It is yields a modified Uzawa algorithm (see p222 of Braess [136] [43]):

Uzawa algorithm (2): assume $\vec{\mathcal{P}}_0$ known. Solve $\mathbb{K} \cdot \vec{\mathcal{V}}_0 = \vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_0$. For $k = 0, 1, 2, ...$, compute

$$\vec{\mathcal{R}}_k = \vec{q}_k = \vec{h} - \mathbb{G}^T \cdot \vec{\mathcal{V}}_k + \mathbb{C} \cdot \vec{\mathcal{P}}_k \tag{10.194}$$

$$\vec{p}_k = \mathbb{G} \cdot q_k \tag{10.195}$$

$$\vec{H}_k = \mathbb{K}^{-1} \cdot \vec{p}_k \tag{10.196}$$

$$\alpha_k = \frac{\vec{q}_k \cdot \vec{q}_k}{\vec{p}_k \cdot \vec{H}_k} \tag{10.197}$$

$$\vec{\mathcal{P}}_k = \vec{\mathcal{P}}_{k-1} - \alpha_k \vec{q}_k \tag{10.198}$$

$$\vec{\mathcal{V}}_k = \vec{\mathcal{V}}_{k-1} + \alpha_k \vec{H}_k \tag{10.199}$$

📚 Relevant Literature: Cahouet & Chabard (1988) [217], Cao (2003) [223].

10.13.3 Conjugate gradient and the Schur complement approach

Since the Schur matrix $\mathbb{S}$ is Symmetric Positive Definite, the Conjugate Gradient (CG) method[44] [600] is very appropriate to solve this system.

Indeed, looking at the definition of Wikipedia: "*In mathematics, the conjugate gradient method is an algorithm for the numerical solution of particular systems of linear equations, namely those whose matrix is symmetric and positive-definite. The conjugate gradient method is often implemented as an iterative algorithm, applicable to sparse systems that are too large to be handled by a direct implementation or other direct methods such as the Cholesky decomposition. Large sparse systems often arise when numerically solving partial differential equations or optimization problems.*"

A simple Google search tells us that the Conjugate Gradient algorithm is as follows:

Algorithm as obtained from Wikipedia.

$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$
if $\mathbf{r}_0$ is sufficiently small, then return $\mathbf{x}_0$ as the result
$\mathbf{p}_0 := \mathbf{r}_0$
$k := 0$
repeat
$\quad \alpha_k := \dfrac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k}$
$\quad \mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$
$\quad \mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A} \mathbf{p}_k$
$\quad$ if $\mathbf{r}_{k+1}$ is sufficiently small, then exit loop
$\quad \beta_k := \dfrac{\mathbf{r}_{k+1}^\top \mathbf{r}_{k+1}}{\mathbf{r}_k^\top \mathbf{r}_k}$
$\quad \mathbf{p}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$
$\quad k := k + 1$
end repeat
return $\mathbf{x}_{k+1}$ as the result

The same algorithm with our notations:
$\vec{r}_0 = \vec{f} - \mathbb{S} \cdot \vec{\mathcal{P}}_0$
$\vec{p}_0 = \vec{r}_0$
$k = 0$
repeat
$\alpha_k = (\vec{r}_k^T \cdot \vec{r}_k) / (\vec{p}_k^T \cdot \mathbb{S} \cdot \vec{p}_k)$
$\vec{\mathcal{P}}_{k+1} = \vec{\mathcal{P}}_k + \alpha_k \vec{p}_k$
$\vec{r}_{k+1} = \vec{r}_k - \alpha_k \mathbb{S} \cdot \vec{p}_k$
$\beta_k = (\vec{r}_{k+1}^T \cdot \vec{r}_{k+1}) / (\vec{r}_k^T \cdot \vec{r}_k)$
$\vec{p}_{k+1} = \vec{r}_{k+1} + \beta_k \vec{p}_k$
$k = k + 1$
end repeat
return $\vec{\mathcal{P}}_{k+1}$ as the result

---

[43] I have slightly altered the indices of the velocities wrt the book
[44] https://en.wikipedia.org/wiki/Conjugate_gradient_method

This algorithm is of course explained in detail in many textbooks such as Saad [1166], in Zhong, Yuen, Moresi & Knepley (2012) [1524], and in Section 16.28.

Let us look at this algorithm more closely. The parts which may prove to be somewhat tricky are those involving the matrix the Schur complement matrix since we wish never to build it explicitly. We start the iterations with a guess pressure $\vec{\mathcal{P}}_0$ (and an initial guess velocity which could be obtained by solving $\mathbb{K} \cdot \vec{\mathcal{V}}_0 = \vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_0$).

$$
\begin{aligned}
\vec{r}_0 &= \underline{\vec{f}} - \mathbb{S} \cdot \vec{\mathcal{P}}_0 & (10.200)\\
&= \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \vec{f} - \vec{h} - (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot \vec{\mathcal{P}}_0 & (10.201)\\
&= \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot (\vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_0) - \vec{h} & (10.202)\\
&= \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{K} \cdot \vec{\mathcal{V}}_0 - \mathbb{C} \cdot \vec{\mathcal{P}}_0 - \vec{h} & (10.203)\\
&= \mathbb{G}^T \cdot \vec{\mathcal{V}}_0 - \mathbb{C} \cdot \vec{\mathcal{P}}_0 - \vec{h} & (10.204)
\end{aligned}
$$

We see that we were able to compute $\mathbb{S} \cdot \vec{\mathcal{P}}_0$ without ever forming the Schur complement matrix explicitely. We now turn to the α_k coefficient:

$$
\alpha_k = \frac{\vec{r}_k^T \cdot \vec{r}_k}{\vec{p}_k \cdot \mathbb{S} \cdot \vec{p}_k} = \frac{\vec{r}_k^T \cdot \vec{r}_k}{\vec{p}_k \cdot (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} + \mathbb{C}) \cdot \vec{p}_k} = \frac{\vec{r}_k^T \cdot \vec{r}_k}{(\mathbb{G} \cdot \vec{p}_k)^T \cdot \mathbb{K}^{-1} \cdot (\mathbb{G} \cdot \vec{p}_k) + \vec{p}_k \cdot \mathbb{C} \cdot \vec{p}_k}
$$

We then define $\tilde{\vec{p}}_k = \mathbb{G} \cdot \vec{p}_k$, so that α_k can be computed as follows:

1. compute $\tilde{\vec{p}}_k = \mathbb{G} \cdot \vec{p}_k$

2. solve $\mathbb{K} \cdot \vec{d}_k = \tilde{\vec{p}}_k$

3. compute

$$
\alpha_k = \frac{\vec{r}_k^T \cdot \vec{r}_k}{\tilde{\vec{p}}_k^T \cdot \vec{d}_k + \vec{p}_k \cdot^T \mathbb{C} \cdot \vec{p}_k}
$$

Then we need to look at the term $\mathbb{S} \cdot \vec{p}_k$:

$$
\mathbb{S} \cdot \vec{p}_k = (\mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G} \cdot + \mathbb{C}) \vec{p}_k = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \tilde{\vec{p}}_k + \mathbb{C} \cdot \vec{p}_k = \mathbb{G}^T \cdot \vec{d}_k + \mathbb{C} \cdot \vec{p}_k
$$

We can then rewrite the CG algorithm as follows:

- choose $\vec{\mathcal{P}}_0$

- compute $\vec{\mathcal{V}}_0$ solution of $\mathbb{K} \cdot \vec{\mathcal{V}}_0 = \vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_0$

- $\vec{r}_0 = \mathbb{G}^T \cdot \vec{\mathcal{V}}_0 - \mathbb{C} \cdot \vec{\mathcal{P}}_0 - \vec{h}$

- if $\vec{r}_0$ is sufficiently small, then return $(\vec{\mathcal{V}}_0, \vec{\mathcal{P}}_0)$ as the result

- $\vec{p}_0 = \vec{r}_0$

- $k = 0$

- repeat

    - compute $\tilde{\vec{p}}_k = \mathbb{G} \cdot \vec{p}_k$
    - solve $\mathbb{K} \cdot \vec{d}_k = \tilde{\vec{p}}_k$
    - compute $\alpha_k = (\vec{r}_k^T \cdot \vec{r}_k)/(\tilde{\vec{p}}_k^T \cdot \vec{d}_k + \vec{p}_k^T \cdot \mathbb{C} \cdot \vec{p}_k)$
    - $\vec{\mathcal{P}}_{k+1} = \vec{\mathcal{P}}_k + \alpha_k \vec{p}_k$

- $\vec{r}_{k+1} = \vec{r}_k - \alpha_k(\mathbb{G}^T \cdot \vec{d}_k + \mathbb{C} \cdot \vec{p}_k)$
- if $\vec{r}_{k+1}$ is sufficiently small, then exit loop
- $\beta_k = (\vec{r}_{k+1}^T \cdot \vec{r}_{k+1})/(\vec{r}_k^T \cdot \vec{r}_k)$
- $\vec{p}_{k+1} = \vec{r}_{k+1} + \beta_k \vec{p}_k$
- $k = k + 1$

- return $\vec{\mathcal{P}}_{k+1}$ as result

We see that we have managed to solve the Schur complement equation with the Conjugate Gradient method without ever building the matrix $\mathbb{S}$. Having obtained the pressure solution $\vec{\mathcal{P}}_{k+1}$, we can easily recover the corresponding velocity with $\mathbb{K} \cdot \vec{\mathcal{V}}_{k+1} = \vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_{k+1}$. However, this is rather unfortunate because it requires yet another solve with the $\mathbb{K}$ matrix. As it turns out, we can slightly alter the above algorithm to have it update the velocity as well so that this last solve is unnecessary.

We have

$$
\begin{aligned}
\vec{\mathcal{V}}_{k+1} &= \mathbb{K}^{-1} \cdot (f - \mathbb{G} \cdot \vec{\mathcal{P}}_{p+1}) & (10.205)\\
&= \mathbb{K}^{-1} \cdot (f - \mathbb{G} \cdot (\vec{\mathcal{P}}_k + \alpha_k \vec{p}_k)) & (10.206)\\
&= \mathbb{K}^{-1} \cdot (f - \mathbb{G} \cdot \vec{\mathcal{P}}_k) - \alpha_k \mathbb{K}^{-1} \cdot \mathbb{G} \cdot \vec{p}_k & (10.207)\\
&= \vec{\mathcal{V}}_k - \alpha_k \mathbb{K}^{-1} \cdot \tilde{\vec{p}}_k & (10.208)\\
&= \vec{\mathcal{V}}_k - \alpha_k \vec{d}_k & (10.209)
\end{aligned}
$$

and we can insert this minor extra calculation inside the algorithm and get the velocity solution nearly for free. The final CG algorithm is then

solver_cg: assume $\vec{\mathcal{P}}_0$ known

- compute $\vec{\mathcal{V}}_0 = \mathbb{K}^{-1} \cdot (\vec{f} - \mathbb{G} \cdot \vec{\mathcal{P}}_0)$

- $\vec{r}_0 = \mathbb{G}^T \cdot \vec{\mathcal{V}}_0 - \mathbb{C} \cdot \vec{\mathcal{P}}_0 - \vec{h}$

- if $\vec{r}_0$ is sufficiently small, then return $(\vec{\mathcal{V}}_0, \vec{\mathcal{P}}_0)$ as the result

- $\vec{p}_0 = \vec{r}_0$

- $k = 0$

- repeat

  - compute $\tilde{\vec{p}}_k = \mathbb{G} \cdot \vec{p}_k$
  - solve $\mathbb{K} \cdot \vec{d}_k = \tilde{p}_k$
  - compute $\alpha_k = (\vec{r}_k^T \cdot \vec{r}_k)/(\tilde{\vec{p}}_k^T \cdot \vec{d}_k + \vec{p}_k^T \cdot \mathbb{C} \cdot \vec{p}_k)$
  - $\vec{\mathcal{P}}_{k+1} = \vec{\mathcal{P}}_k + \alpha_k \vec{p}_k$
  - $\vec{\mathcal{V}}_{k+1} = \vec{\mathcal{V}}_k - \alpha_k \vec{d}_k$
  - $\vec{r}_{k+1} = \vec{r}_k - \alpha_k(\mathbb{G}^T \cdot \vec{d}_k + \mathbb{C} \cdot \vec{p}_k)$
  - if $\vec{r}_{k+1}$ is sufficiently small ($||\vec{r}_{k+1}||_2/||\vec{r}_0||_2 < tol$), then exit loop
  - $\beta_k = (r_{k+1}^T r_{k+1})/(r_k^T r_k)$
  - $\vec{p}_{k+1} = \vec{r}_{k+1} + \beta_k \vec{p}_k$

- $k = k + 1$

- return $\vec{\mathcal{P}}_{k+1}$ as result

Remark. *The matrix $\mathbb{C}$ is rarely present unless for example when stabilised elements are used such as the stabilised $Q_1 \times P_0$ or the stabilised $Q_1 \times Q_1$ elements.*

This iterative algorithm will converge to the solution with a rate which depends on the condition number of the $\mathbb{S}$ matrix, which is not easy to compute since $\mathbb{S}$ is never built. However, it has been established that large viscosity contrasts in the domain will have a negative impact on the convergence.

Remark. *This algorithm requires one solve with matrix $\mathbb{K}$ per iteration but says nothing about the method employed to do so (direct or iterative solver) nor the corresponding preconditioner.*

One thing we know improves the convergence of any iterative solver is the use of a preconditioner matrix and therefore now focus on the Preconditioned Conjugate Gradient (PCG) method. Once again we turn to Wikipedia[45]:

Algorithm as obtained from Wikipedia.

$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{z}_0 := \mathbf{M}^{-1}\mathbf{r}_0$
$\mathbf{p}_0 := \mathbf{z}_0$
$k := 0$
repeat

$\qquad \alpha_k := \dfrac{\mathbf{r}_k^\mathsf{T}\mathbf{z}_k}{\mathbf{p}_k^\mathsf{T}\mathbf{A}\mathbf{p}_k}$

$\qquad \mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{p}_k$
$\qquad \mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{p}_k$
$\qquad$ if $\mathbf{r}_{k+1}$ is sufficiently small **then** exit loop **end if**
$\qquad \mathbf{z}_{k+1} := \mathbf{M}^{-1}\mathbf{r}_{k+1}$

$\qquad \beta_k := \dfrac{\mathbf{z}_{k+1}^\mathsf{T}\mathbf{r}_{k+1}}{\mathbf{z}_k^\mathsf{T}\mathbf{r}_k}$

$\qquad \mathbf{p}_{k+1} := \mathbf{z}_{k+1} + \beta_k\mathbf{p}_k$
$\qquad k := k + 1$
end repeat
The result is $\mathbf{x}_{k+1}$

The same algorithm with our notations:
$\vec{r}_0 = \vec{f} - \mathbb{S} \cdot \vec{\mathcal{P}}_0$
$\vec{z}_0 = \boldsymbol{M}^{-1} \cdot \vec{r}_0$
$\vec{p}_0 = \vec{z}_0$
$k = 0$
repeat
$\alpha_k = (\vec{r}_k^T \cdot \vec{z}_k)/(\vec{p}_k^T \cdot \mathbb{S} \cdot \vec{p}_k)$
$\vec{\mathcal{P}}_{k+1} = \vec{\mathcal{P}}_k + \alpha_k \vec{p}_k$
$\vec{r}_{k+1} = \vec{r}_k - \alpha_k \, \mathbb{S} \cdot \vec{p}_k$
$\vec{z}_{k+1} = \boldsymbol{M}^{-1} \cdot \vec{r}_{k+1}$
$\beta_k = (\vec{z}_{k+1}^T \cdot \vec{r}_{k+1})/(\vec{z}_k^T \cdot \vec{r}_k)$
$\vec{p}_{k+1} = \vec{z}_{k+1} + \beta_k \vec{p}_k$
$k = k + 1$
end repeat
return $\vec{\mathcal{P}}_{k+1}$ as the result

Unsurprisingly we find the same algorithm in Saad [1166]:

ALGORITHM **9.1** *Preconditioned Conjugate Gradient*

1. $\quad$ Compute $r_0 := b - Ax_0$, $z_0 = M^{-1}r_0$, and $p_0 := z_0$
2. $\quad$ For $j = 0, 1, \ldots$, until convergence Do:
3. $\qquad \alpha_j := (r_j, z_j)/(Ap_j, p_j)$
4. $\qquad x_{j+1} := x_j + \alpha_j p_j$
5. $\qquad r_{j+1} := r_j - \alpha_j Ap_j$
6. $\qquad z_{j+1} := M^{-1}r_{j+1}$
7. $\qquad \beta_j := (r_{j+1}, z_{j+1})/(r_j, z_j)$
8. $\qquad p_{j+1} := z_{j+1} + \beta_j p_j$
9. $\quad$ EndDo

Note that in the algorithm above the preconditioner matrix $\boldsymbol{M}$ has to be symmetric positive-definite and fixed, i.e., cannot change from iteration to iteration. We see that this algorithm introduces an additional vector $\vec{z}$ and a solve with the matrix $\boldsymbol{M}$ at each iteration, which means that $\boldsymbol{M}$ must be such that solving $\boldsymbol{M} \cdot \vec{x} = \vec{f}$ where $\vec{f}$ is a given rhs vector must be cheap. Ultimately, the PCG algorithm applied to the Schur complement equation takes the form:

---

[45]https://en.wikipedia.org/wiki/Conjugate_gradient_method

Following Zhong *et al.* [1524] one can define the following matrix as preconditioner:

$$\boldsymbol{M} = diag\left[\mathbb{G}^T (diag[\mathbb{K}])^{-1}\mathbb{G}\right]$$

which is the preconditioner used for the Citcom codes (see appendix **??**). It can be constructed while the FEM matrix is being built/assembled and it is trivial to invert. The entries in $diag[\mathbb{K}]$ are the average viscosity in the elements associated with a given degree of freedom.

Another very cheap way of building $\boldsymbol{M}$ for $Q_1 \times P_0$ lements is to realise that the matrix $\mathbb{S}$ has dimensions element surface/volume divided by viscosity. We can then postulate

$$M_{e,e} = \frac{|\Omega|_e}{\eta_e}$$

where e is an element and η_e is the (average viscosity) inside the element. For higher order elements, we need to use the pressure mass matrix.

These two preconditioners and two other variants are implemented in STONE 16 for $Q_1 \times P_0$ elements.

10.13.4 Generalized Conjugate Residual approach (Geenen *et al.* (2009))

This approach is presented in Geenen *et al.* (2009) [471]. The saddle point problem arising from the constrained Stokes equation is solved with a Krylov method, GCR [1429], right preconditioned

(postconditioned) with a block triangular preconditioner (BTR) [140].

The preconditioner $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ 0 & -\tilde{\mathbb{S}} \end{array} \right)$$

The GCR algorithm [388] in this case is taken from Vuik *et al.* (2000) [1430] and makes use of the block triangular preconditioner as follows:

$\vec{r}_0 = \vec{b} - \boldsymbol{A} \cdot \vec{x}^0$

for k=0,1,2,...

- $\vec{s}^{k+1} = \boldsymbol{P}^{-1} \cdot \vec{r}^k$
- $\vec{v}^{k+1} = \boldsymbol{A} \cdot \vec{s}^{k+1}$
- for i=0,1,...k
    * $\vec{v}^{k+1} = \vec{v}^{k+1} - (\vec{v}^{k+1}, \vec{v}^i)\vec{v}^i$
    * $\vec{s}^{k+1} = \vec{s}^{k+1} - (\vec{v}^{k+1}, \vec{v}^i)\vec{s}^i$
- end for
- $\vec{v}^{k+1} = \vec{v}^{k+1}/\|\vec{v}^{k+1}\|_2$
- $\vec{s}^{k+1} = \vec{s}^{k+1}/\|\vec{v}^{k+1}\|_2$
- $\vec{x}^{k+1} = \vec{x}^k + (\vec{v}^{k+1}, \vec{r}^k)\vec{s}^{k+1}$
- $\vec{r}^{k+1} = \vec{r}^k - (\vec{v}^{k+1}, \vec{r}^k)\vec{v}^{k+1}$

end for

As explained in Geenen *et al.* , instead of constructing $\boldsymbol{P}^{-1}$ explicitely and applying it to $\vec{r}$, we instead solve the system $\boldsymbol{P} \cdot \vec{s} = \vec{r}$. We first decompose $\vec{r}$ and $\vec{s}$ as follows:

$$\vec{r}^k = \left(\begin{array}{c} \vec{r}_\mathcal{V}^k \\ \vec{r}_p^k \end{array} \right) \qquad \vec{s}^{k+1} = \left(\begin{array}{c} \vec{s}_\mathcal{V}^{k+1} \\ \vec{s}_p^{k+1} \end{array} \right)$$

so that we have to solve

$$\left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ 0 & -\tilde{\mathbb{S}} \end{array} \right) \cdot \left(\begin{array}{c} \vec{s}_\mathcal{V}^{k+1} \\ \vec{s}_p^{k+1} \end{array} \right) = \left(\begin{array}{c} \vec{r}_\mathcal{V}^k \\ \vec{r}_p^k \end{array} \right)$$

This is actually rather trivial because of the upper triangular nature of the preconditioner $\boldsymbol{P}$. It immediately follows:

$$\tilde{\mathbb{S}} \cdot \vec{s}_p^{k+1} = -\vec{r}_p^k \tag{10.210}$$
$$\mathbb{K} \cdot \vec{s}_\mathcal{V}^{k+1} = \vec{r}_\mathcal{V}^k - \mathbb{G} \cdot \vec{s}_p^{k+1} \tag{10.211}$$

As before we now must specify how we solve the above two equations (and we must therefore make a choice about the approximate Schur complement $\tilde{\mathbb{S}}$).

In the paper they take $\boldsymbol{M}_p$, the pressure mass matrix scaled with the inverse of viscosity as an approximation to the Schur complement $\tilde{\mathbb{S}}$, which is spectrally equivalent. Note that sometimes this mass matrix can be lumped which makes solving with it trivial and fast.

The inner solve with $\mathbb{K}$ is carried out with a CG solvers preconditioned with AMG. They state that "Using AMG as a preconditioner to CG for the subsystem solution guarantees h-independent convergence of the solver during the preconditioner construction phase."

10.13.5 The Arrow-Hurwicz algorithm

This is borrowed from algorithm 8.7 in Saad [1166].

Select an initial guess $\vec{\mathcal{V}}^0$ and $\vec{\mathcal{P}}^0$

for k=0,1,2,... until convergence

– $\vec{\mathcal{V}}^{k+1} = \vec{\mathcal{V}}^k + \epsilon(\vec{f} - \mathbb{K} \cdot \vec{\mathcal{V}}^k - \mathbb{G} \cdot \vec{\mathcal{P}}^k)$

– $\vec{\mathcal{P}}^{k+1} = \vec{\mathcal{P}}^k + \omega(\mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} - \vec{h})$

The above algorithm is a block-iteration of the form

$$\left(\begin{array}{cc} I & 0 \\ -\omega\mathbb{G}^T & I \end{array} \right) \cdot \left(\begin{array}{c} \vec{\mathcal{V}}^{k+1} \\ \vec{\mathcal{P}}^{k+1} \end{array} \right) = \left(\begin{array}{cc} I - \epsilon\mathbb{K} & -\epsilon\mathbb{G} \\ 0 & I \end{array} \right) \cdot \left(\begin{array}{c} \vec{\mathcal{V}}^k \\ \vec{\mathcal{P}}^k \end{array} \right) + \left(\begin{array}{c} \epsilon\vec{f} \\ -\omega\vec{h} \end{array} \right)$$

10.13.6 Using MINRES a la Burstedde *et al.* (2008)

This approach is presented in Burstedde *et al.* (2008) [204]. They state that neglecting the off-diagonal blocks motivates use of the symmetric positive definite preconditioner:

$$\boldsymbol{P} = \left(\begin{array}{cc} \tilde{\mathbb{K}} & 0 \\ 0 & \tilde{\mathbb{S}} \end{array} \right)$$

where $\tilde{\mathbb{K}}$ is a variable-viscosity discrete vector Laplacian approximation of $\mathbb{K}$ (see explanations in [202]), which is motivated by the fact that for constant viscosity and Dirichlet boundary conditions, $\mathbb{K}$ and $\tilde{\mathbb{K}}$ are equivalent. $\tilde{\mathbb{S}}$ is an approximation of the Schur complement given by a lumped mass matrix weighted by the inverse viscosity η^{-1}. The resulting diagonal matrix $\tilde{\mathbb{S}}$ is spectrally equivalent to $\mathbb{S}$ [394]. They also use AMG as preconditioner for the inner solves.

Note that Burstedde *et al.* (2008) [204] relies on stabilised $Q_1 \times Q_1$ elements from Dohrmann & Bochev [358] so that their Stokes matrix does feature the associated $-\mathbb{C}$ block. Subsequent papers do so too, see Burstedde *et al.* (2009) [202], Burstedde *et al.* (2013) [203]. The same solver structure based on MINRES is used in these articles too.

10.13.7 The Augmented Lagrangian approach

see LaCoDe paper [341].

We start from the saddle point Stokes system:

$$\left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{array} \right) \cdot \left(\begin{array}{c} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{array} \right) = \left(\begin{array}{c} \vec{f} \\ \vec{h} \end{array} \right) \tag{10.212}$$

The AL method consists of subtracting $\lambda^{-1}\mathbb{M}_p \cdot \vec{\mathcal{P}}$ from the left and right-side of the mass conservation equation (where $\mathbb{M}_p$ is the pressure mass matrix) and introducing the following iterative scheme:

$$\left(\begin{array}{cc} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\lambda^{-1}\mathbb{M}_p \end{array} \right) \cdot \left(\begin{array}{c} \vec{\mathcal{V}}^{k+1} \\ \vec{\mathcal{P}}^{k+1} \end{array} \right) = \left(\begin{array}{c} \vec{f} \\ \vec{h} - \lambda^{-1}\mathbb{M}_p \cdot \vec{\mathcal{P}}^k \end{array} \right) \tag{10.213}$$

where k is the iteration counter and λ is an artificial compressibility term which has the dimensions of dynamic viscosity. The choice of λ can be difficult as too low or too high a value yields either erroneous results and/or terribly ill-conditioned matrices. LaCoDe paper (!!) use such a method

and report that $\lambda = \max_\Omega(\eta)$ works well. Note that at convergence we have $||\vec{\mathcal{P}}^{k+1} - \vec{\mathcal{P}}^k|| < \epsilon$ and then Eq.(10.213) converges to Eq.(10.212) and the velocity and pressure fields are solution of the unmodified system Eq.(10.212).

The introduction of this term serves one purpose: allowing us to solve the system in a segregated manner (i.e. computing successive iterates of the velocity and pressure fields until convergence is reached). The second line of Eq. (10.213) is

$$\mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} - \lambda^{-1}\mathbb{M}_p \cdot \vec{\mathcal{P}}^{k+1} = \vec{h} - \lambda^{-1}\mathbb{M}_p \cdot \vec{\mathcal{P}}^k$$

and can therefore be rewritten

$$\vec{\mathcal{P}}^{k+1} = \vec{\mathcal{P}}^k + \lambda\mathbb{M}_p^{-1} \cdot (\mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} - \vec{h})$$

We can then substitute this expression of $\vec{\mathcal{P}}^{k+1}$ in the first equation. This yields:

$$\mathbb{K} \cdot \vec{\mathcal{V}}^{k+1} = \vec{f} - \mathbb{G} \cdot \mathcal{P}^{k+1}) \tag{10.214}$$

$$\mathbb{K} \cdot \vec{\mathcal{V}}^{k+1} = \vec{f} - \mathbb{G} \cdot (\vec{\mathcal{P}}^k + \lambda\mathbb{M}_p^{-1} \cdot (\mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} - \vec{h})) \tag{10.215}$$

$$\mathbb{K} \cdot \vec{\mathcal{V}}^{k+1} + \lambda\mathbb{G} \cdot \mathbb{M}_p^{-1} \cdot \mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} = \vec{f} - \mathbb{G} \cdot (\vec{\mathcal{P}}^k - \lambda\mathbb{M}_p^{-1}\vec{h})) \tag{10.216}$$

$$\underbrace{\left(\mathbb{K} + \lambda\mathbb{G} \cdot \mathbb{M}_p^{-1} \cdot \mathbb{G}^T\right)}_{\tilde{\mathbb{K}}} \cdot \vec{\mathcal{V}}^{k+1} = \underbrace{\vec{f} - \mathbb{G} \cdot (\vec{\mathcal{P}}^k - \lambda\mathbb{M}_p^{-1}\vec{h}))}_{\vec{f}^{k+1}} \tag{10.217}$$

$$\tag{10.218}$$

The iterative algorithm goes as follows:

1. if it is the first timestep, set $\vec{\mathcal{P}}^0 = 0$, otherwise set it to the pressure of the previous timestep.

2. calculate $\tilde{\mathbb{K}}$

3. calculate $\vec{f}^{k+1}$

4. solve $\tilde{\mathbb{K}} \cdot \vec{\mathcal{V}}^{k+1} = \vec{f}^{k+1}$

5. update pressure with $\vec{\mathcal{P}}^{k+1} = \vec{\mathcal{P}}^k + \lambda\mathbb{M}_p^{-1} \cdot (\mathbb{G}^T \cdot \vec{\mathcal{V}}^{k+1} - \vec{h})$

Remark. *If discontinuous pressures are used, the pressure mass matrix can be inverted element by element which is cheaper than inverting $\mathbb{M}_p$ as a whole.*

Remark. *This method has obvious ties with the penalty method.*

Remark. *If $\lambda >> \max_\Omega \eta$ then the matrix $\tilde{\mathbb{K}}$ is ill-conditioned and an iterative solver must be used.*

10.13.8   The SIMPLE method

What follows is borrowed from Volker John. *Finite Element Methods for Incompressible Flow Problems.* Springer, 2016. ISBN: 978-3-319-45749-9. DOI: 10.1007/978-3-319-45750-5, page 666.

The SIMPLE method (Semi-Implicit Method for Pressure-Linked Equations) has been introduced by Patankar & Spalding (1972) [1041] as an iterative method to solve the finite volume discretized incompressible Navier-Stokes equations.

The algorithm is based on the following steps:

- First the pressure is assumed to be known from the previous iteration.

- Then the velocity is solved from the momentum equations. The newly obtained velocities do not satisfy the continuity equation since the pressure is only a guess.

- In the next substeps the velocities and pressures are corrected in order to satisfy the discrete continuity equation.

SIMPLE relies on the block LU decomposition

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\mathbb{S} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & \mathbb{K}^{-1} \cdot \mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix}
\tag{10.219}
$$

The approximation $\mathbb{K}^{-1}$ as $\boldsymbol{D}_{\mathbb{K}}^{-1} = (diag(\mathbb{K}))^{-1}$ leads to the SIMPLE algorithm. In this case the approximation of the Schur complement matrix is given by $\tilde{\mathbb{S}} = \mathbb{G}^T \cdot \boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G}$ and the decomposition looks like

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \simeq \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\tilde{\mathbb{S}} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & \boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix}
$$

Thus one iteration of SIMPLE solves the following system:

$$
\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{C} \end{pmatrix} \simeq \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\tilde{\mathbb{S}} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & \boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix}
$$

Before we can write out the SIMPLE algorithm, we must first take a small detour via so-called distributive iterative methods [1430, 659]. Let us consider the linear system

$$
\boldsymbol{A} \cdot \vec{x} = \vec{b}
$$

A stationary iterative method is defined as follows:

$$
\vec{x}^{k+1} = \boldsymbol{B} \cdot \vec{x}^k + \vec{c}
$$

where $\vec{c} = (\boldsymbol{I} - \boldsymbol{B}) \cdot \boldsymbol{A}^{-1} \cdot \vec{b}$. Left-multiplying all terms by $(\boldsymbol{I} - \boldsymbol{B})^{-1}$ first and then left-multiplying again by $\boldsymbol{A}$ we arrive at:

$$
\boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1} \cdot \vec{x}^{k+1} = \boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1} \cdot \boldsymbol{B} \cdot \vec{x}^k + \boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1} \cdot \vec{c}
$$

We define $\boldsymbol{M} = \boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1}$ so that now

$$
\boldsymbol{M} \cdot \vec{x}^{k+1} = \boldsymbol{M} \cdot \boldsymbol{B} \cdot \vec{x}^k + \vec{b}
$$

We define $\boldsymbol{N} = \boldsymbol{M} \cdot \boldsymbol{B}$ and finally

$$
\boldsymbol{M} \cdot \vec{x}^{k+1} = \boldsymbol{N} \cdot \vec{x}^k + \vec{b}
$$

Note that $\boldsymbol{M} - \boldsymbol{N} = \boldsymbol{M} - \boldsymbol{M} \cdot \boldsymbol{B} = \boldsymbol{M} \cdot (\boldsymbol{I} - \boldsymbol{B}) = \boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1} \cdot (\boldsymbol{I} - \boldsymbol{B}) = \boldsymbol{A}$. Let us now write the original system $\boldsymbol{A} \cdot \vec{x} = \vec{b}$ as $(\boldsymbol{A} \cdot \boldsymbol{B}) \cdot (\boldsymbol{B}^{-1} \cdot \vec{x}) = \vec{b}$ or, $\underline{\boldsymbol{A}} \cdot \vec{\underline{x}} = \vec{b}$ with $\vec{x} = \boldsymbol{B} \cdot \vec{\underline{x}}$ and $\underline{\boldsymbol{A}} = \boldsymbol{A} \cdot \boldsymbol{B}$. Splitting $\underline{\boldsymbol{A}} = \boldsymbol{M} - \boldsymbol{N}$ again yields

$$\boldsymbol{M} \cdot \vec{\underline{x}}^{k+1} = \boldsymbol{N} \cdot \vec{\underline{x}}^{k} + \vec{b}$$

Using $\vec{x} = \boldsymbol{B} \cdot \vec{\underline{x}}$, we get

$$\boldsymbol{M} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k+1} = \boldsymbol{N} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} + \vec{b}$$

We then have

$$
\begin{aligned}
\vec{x}^{k+1} &= \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [\boldsymbol{N} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} + \vec{b}] & (10.220) \\
&= \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [(\boldsymbol{M} - \underline{\boldsymbol{A}}) \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} + \vec{b}] & (10.221) \\
&= \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [(\boldsymbol{M} - \boldsymbol{A} \cdot \boldsymbol{B}) \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} + \vec{b}] & (10.222) \\
&= \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [\boldsymbol{M} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} - \boldsymbol{A} \cdot \boldsymbol{B} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} + \vec{b}] & (10.223) \\
&= \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [\boldsymbol{M} \cdot \boldsymbol{B}^{-1} \cdot \vec{x}^{k} - \boldsymbol{A} \cdot \vec{x}^{k} + \vec{b}] & (10.224) \\
&= \vec{x}^{k} + \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot [\vec{b} - \boldsymbol{A} \cdot \vec{x}^{k}] & (10.225)
\end{aligned}
$$

Finally, we have the following recursion:

$$\boxed{\vec{x}^{k+1} = \vec{x}^{k} + \boldsymbol{B} \cdot \boldsymbol{M}^{-1} \cdot (\vec{b} - \boldsymbol{A} \cdot \vec{x}^{k})} \qquad (10.226)$$

Coming back to the SIMPLE algorithm, we start from

$$\boldsymbol{A} = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix}$$

The matrix $\boldsymbol{B}$ is then chosen to be

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I} & -\mathbb{K}^{-1}\mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix}$$

We then have

$$\boldsymbol{A} \cdot \boldsymbol{B} = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & -\mathbb{K}^{-1}\mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix} = \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\mathbb{S} \end{pmatrix}$$

where $\mathbb{S} = \mathbb{G}^T \cdot \mathbb{K}^{-1} \cdot \mathbb{G}$. Let us recall that we define $\boldsymbol{D}_\mathbb{K} = diag(\mathbb{K})$ and $\hat{\mathbb{S}} = \mathbb{G}^T \cdot \boldsymbol{D}_\mathbb{K}^{-1} \cdot \mathbb{G}$. We further define

$$\boldsymbol{M} = \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\hat{\mathbb{S}} \end{pmatrix}$$

and $\boldsymbol{N}$ follows from the splitting $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{M} - \boldsymbol{N}$. (Note that we do not need to form nor use $\boldsymbol{N}$).

The standard SIMPLE algorithm also replaces $\mathbb{K}^{-1}$ by $\boldsymbol{D}_\mathbb{K}^{-1}$ in $\boldsymbol{B}$ so that $\boldsymbol{B}$ is approximated by:

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{D}_\mathbb{K}^{-1}\mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix}$$

in the iterations. We can define

$$\vec{r}^{k} = \vec{b} - \boldsymbol{A} \cdot \vec{x}^{k} = \begin{pmatrix} \vec{f} \\ \vec{h} \end{pmatrix} - \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}}^{k} \\ \vec{\mathcal{P}}^{k} \end{pmatrix} = \begin{pmatrix} \vec{r}_\mathcal{V}^{k} \\ \vec{r}_\mathcal{P}^{k} \end{pmatrix}$$

The iteration loop of Eq. (10.226) then takes the form

$$\begin{pmatrix} \vec{\mathcal{V}}^{k+1} \\ \vec{\mathcal{P}}^{k+1} \end{pmatrix} = \begin{pmatrix} V^{k} \\ P^{k} \end{pmatrix} + \boldsymbol{B}\boldsymbol{M}^{-1} \begin{pmatrix} r_V^{k} \\ r_P^{k} \end{pmatrix} = \begin{pmatrix} V^{k} \\ P^{k} \end{pmatrix} + \begin{pmatrix} \delta V^{k} \\ \delta P^{k} \end{pmatrix} \qquad \text{with} \qquad \begin{pmatrix} \delta V^{k} \\ \delta P^{k} \end{pmatrix} = \boldsymbol{B}\boldsymbol{M}^{-1} \begin{pmatrix} \vec{r}_\mathcal{V}^{k} \\ \vec{r}_\mathcal{P}^{k} \end{pmatrix}$$

This last equation can be rewritten[46]:

$$\boldsymbol{M} \cdot \left[\boldsymbol{B}^{-1} \cdot \begin{pmatrix} \delta\vec{\mathcal{V}}^k \\ \delta\vec{\mathcal{P}}^k \end{pmatrix} \right] = \begin{pmatrix} \vec{r}_{\mathcal{V}}^k \\ \vec{r}_{\mathcal{P}}^k \end{pmatrix}$$

We then have to solve

$$\boldsymbol{M} \cdot \begin{pmatrix} \delta^\star\vec{\mathcal{V}}^k \\ \delta^\star\vec{\mathcal{P}}^k \end{pmatrix} = \begin{pmatrix} \mathbb{K} & 0 \\ \mathbb{G}^T & -\hat{\mathbb{S}} \end{pmatrix} \cdot \begin{pmatrix} \delta^\star\vec{\mathcal{V}}^k \\ \delta^\star\vec{\mathcal{P}}^k \end{pmatrix} = \begin{pmatrix} \vec{r}_{\mathcal{V}}^k \\ \vec{r}_{\mathcal{P}}^k \end{pmatrix} \tag{10.227}$$

and then compute

$$\begin{pmatrix} \delta\vec{\mathcal{V}}^k \\ \delta\vec{\mathcal{P}}^k \end{pmatrix} = \boldsymbol{B} \begin{pmatrix} \delta^\star\vec{\mathcal{V}}^k \\ \delta^\star\vec{\mathcal{P}}^k \end{pmatrix} \tag{10.228}$$

Fortunately Eq. (10.227) translates into:

$$\mathbb{K} \cdot \delta^\star\vec{\mathcal{V}}^k = \vec{r}_{\mathcal{V}}^k \tag{10.229}$$

$$\hat{\mathbb{S}} \cdot \delta^\star\vec{\mathcal{P}}^k = -\vec{r}_P^k + \mathbb{G}^T \cdot \delta^\star\vec{V}^k \tag{10.230}$$

and Eq. (10.228) translates into:

$$\begin{pmatrix} \delta\vec{\mathcal{V}}^k \\ \delta\vec{\mathcal{P}}^k \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G} \\ 0 & \boldsymbol{I} \end{pmatrix} \cdot \begin{pmatrix} \delta^\star\vec{\mathcal{V}}^k \\ \delta^*\vec{\mathcal{P}}^k \end{pmatrix}$$

or,

$$\delta\vec{\mathcal{V}}_k = \delta^\star\vec{\mathcal{V}}^k - \boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G} \cdot \delta^\star\vec{\mathcal{P}}_k \tag{10.231}$$

$$\delta\vec{\mathcal{P}}_k = \delta^\star\vec{\mathcal{P}}^k \tag{10.232}$$

The final algorithm will then look as follows:

1. compute the residuals

$$\begin{aligned} \vec{r}_{\mathcal{V}} &= \vec{f} - \mathbb{K} \cdot \vec{\mathcal{V}}^{(k)} - \mathbb{G} \cdot \vec{\mathcal{P}}^{(k)} \\ \vec{r}_{\mathcal{P}} &= \vec{h} - \mathbb{G}^T \cdot \vec{\mathcal{V}}^{(k)} \end{aligned} \tag{10.233}$$

2. Solve $\mathbb{K} \cdot \delta^\star\vec{\mathcal{V}}^k = \vec{r}_{\mathcal{V}}^k$

3. Solve $\hat{\mathbb{S}} \cdot \delta^\star\vec{\mathcal{P}}^k = \vec{r}_{\mathcal{P}}^k - \mathbb{G}^T \cdot \delta^\star V^k$

4. Compute $\delta\vec{\mathcal{V}}^k = \delta^\star\vec{\mathcal{V}}^k - \boldsymbol{D}_{\mathbb{K}}^{-1} \cdot \mathbb{G} \cdot \delta^\star\vec{\mathcal{P}}_k$

5. Update $\delta\vec{\mathcal{P}}^k = \delta^\star\vec{\mathcal{P}}^k$

6. Update

$$\begin{aligned} \vec{\mathcal{V}}^{(k+1)} &= \vec{\mathcal{V}}^{(k)} + \omega_{\mathcal{V}}\delta\vec{\mathcal{V}}^{(k)} \\ \vec{\mathcal{P}}^{(k+1)} &= \vec{\mathcal{P}}^{(k)} + \omega_{\mathcal{P}}\delta\vec{\mathcal{P}}^{(k)} \end{aligned} \tag{10.234}$$

where the parameters $\omega_{\mathcal{V}}$ and $\omega_{\mathcal{P}}$ are between 0 and 1.

Note that SIMPLE can be used as left and as right preconditioner, see page 669 of John [688].

---

[46]Remember that $(\boldsymbol{A} \cdot \boldsymbol{B})^{-1} = \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^{-1}$

Also, John states that: "SIMPLE is easily to implement, which makes it attractive. It relies on the already assembled matrix blocks. Only the approximation $\hat{\mathbb{S}}$ of the Schur complement matrix has to be computed. This matrix couples pressure degrees of freedom that are usually not coupled in finite element approximations of the diffusion operator, but it is still a sparse matrix. The efficiency of SIMPLE depends on how good $\mathbb{K}^{-1}$ is approximated by its diagonal."

📚 Relevant Literature:

- C. Echevarria Serur. "Fast iterative methods for solving the incompressible Navier-Stokes equations". PhD thesis. TU Delft, 2013,

- D. Braess and R. Sarazin. "An Efficient Smoother for the Stokes Problem". In: *Applied Numerical Math.* 23 (1997), pp. 3–20,

- M. ur Rehman, C. Vuik, and G. Segal. "SIMPLE-type preconditioners for the Oseen problem". In: *Int. J. Num. Meth. Fluids* 61 (2009), pp. 432–452. DOI: 10.1002/fld.1957 for SIMPLE(R) algorithm,

- Alik Ismail-Zadeh and Paul Tackley. *Computational Methods for Geodynamics.* Cambridge University Press, 2010

10.13.9 The GMRES approach - NOT FINISHED

The Generalized Minimal Residual method [1167] is an extension of MINRES (which is only applicable to symmetric systems) to unsymmetric systems. Like MINRES, it generates a sequence of orthogonal vectors and combines these through a least-squares solve and update. However, in the absence of symmetry this can no longer be done with short recurrences. As a consequence, all previously computed vectors in the orthogonal sequence have to be retained and for this reason "restarted" versions of the method are used.

It must be said that the (preconditioned) GMRES method is actually much more difficult to implement than the (preconditioned) Conjugate Gradient method. However, since it can deal with unsymmetric matrices, it means that it can be applied directly to the Stokes system matrix (as opposed to the CG method which is used on the Schur complement equation).

Relevant Literature: [387, p208] [1166, 1165] [52] [36]

finish GMRES algo description. not sure what to do, hard to explain, not easy to code.

Algorithm 3 GMRESR algorithm

1. x_0 is an initial guess and $r_o = b - Ax_0$
2. For $j = 1, 2, 3...$
3. $s_i = P_{m,i-1}(A)r_{i-1}$,
        s_i be the approximate solution of $As = r_{i-1}$
        obtained after m steps of an iterative method
4. $v_i = As_i$
5.   For $j = 1$ to $i - 1$
6. $\alpha = (v_i, v_j)$,
7. $v_i = v_i - \alpha v_j, s_i = s_i - \alpha s_j$,
8.   End
9. $v_i = v_i/\|v_i\|_2, s_i = s_i/\|v_i\|_2$
10. $x_i = x_{i-1} + (r_{i-1}, v_i)s_i$;
11. $r_i = r_{i-1} - (r_{i-1}, v_i)v_i$;
12. End

Taken from ur Rehman, vuik & Segal.

the FGMRES approach [351]

Relevant Literature[1029, 895, 457, 759, 760, 764, 1160]

10.14 Boundary conditions

10.14.1 Imposing Dirichlet boundary conditions

Let us consider a quadrilateral element with one degree of freedom per node and let us assume that we are solving the temperature equation. The local matrix and right-hand side vector are given by

$$A_{el}(4 \times 4) \qquad and \qquad B_{el}(4)$$

Let us assume that we want to impose $\tilde{T} = 10$ on the third node (local coordinates numbering). For instance, having built A_{el} and B_{el}, the system looks like :

$$\begin{pmatrix} 3 & 1 & 6 & 9 \\ 5 & 2 & 2 & 8 \\ 7 & 4 & 11 & 2 \\ 9 & 6 & 4 & 3 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 1 \\ 2 \end{pmatrix}$$

which can be rewritten

$$3T_1 + T_2 + 6T_3 + 9T_4 = 4$$

$$5T_1 + 2T_2 + 2T_3 + 8T_4 = 3$$

$$7T_1 + 4T_2 + 11T_3 + 2T_4 = 1$$

$$9T_1 + 6T_2 + 4T_3 + 3T_4 = 2$$

or,

$$3T_1 + T_2 + \quad + 9T_4 = 4 - 6T_3$$

$$5T_1 + 2T_2 + \quad + 8T_4 = 3 - 2T_3$$

$$7T_1 + 4T_2 + 11T_3 + 2T_4 = 1$$

$$9T_1 + 6T_2 + \quad + 3T_4 = 2 - 4T_3$$

- <u>Technique 1</u>: Replace the hereabove system by

$$\begin{pmatrix} 3 & 1 & 6 & 9 \\ 5 & 2 & 2 & 8 \\ 7 & 4 & 11 + 10^{12} & 2 \\ 9 & 6 & 4 & 3 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ \tilde{T} \times (11 + 10^{12}) \\ 2 \end{pmatrix}$$

- <u>Technique 2</u>: One can choose not to solve for T_3 anymore, i.e. not to consider it as a degree of freedom and therefore write:

$$3T_1 + T_2 + 9T_4 = 4 - 6T_3$$

$$5T_1 + 2T_2 + 8T_4 = 3 - 2T_3$$

$$9T_1 + 6T_2 + 3T_4 = 2 - 4T_3$$

- <u>Technique 3</u>: Since we want to impose $T_3 = 10$, then we can write

$$3T_1 + T_2 + \quad +9T_4 = 4 - 6T_3$$

$$5T_1 + 2T_2 + \quad +8T_4 = 3 - 2T_3$$

$$0 + 0 + T_3 + 0 = 10$$

$$9T_1 + 6T_2 + \quad +3T_4 = 2 - 4T_3$$

and in matrix form :

$$
\begin{pmatrix}
3 & 1 & 0 & 9 \\
5 & 2 & 0 & 8 \\
0 & 0 & 1 & 0 \\
9 & 6 & 0 & 3
\end{pmatrix}
\begin{pmatrix}
T_1 \\ T_2 \\ T_3 \\ T_4
\end{pmatrix}
=
\begin{pmatrix}
4 - A_{13}T_3 \\
3 - A_{23}T_3 \\
10 \\
2 - A_{43}T_3
\end{pmatrix}
$$

The first technique is not a good idea in practice as it introduces very large values and will likely derail the solver. The second option is somewhat difficult to implement as it means that elemental matrix and rhs sizes will change from element to element and it therefore requires more book-keeping. The third technique is the one adopted throughout this document.

As shown in Wu, Xu, and Li [1472] (2008), it is better to replace the 1 on the diagonal by the former diagonal term as it reduces the condition number of the matrix. The rhs must then be modified accordingly.

🔖 Relevant LiteratureBehr (2004) [74]

This is an excerpt of an email sent to me by Dave May in May 2014: *Never ever ever impose bc's using a penalty approach. For problems with a fixed mesh topology and time dependent Dirichlet domain (e.g. the segment of the boundary with Dirichlet bc's maybe change size/shape over time - for example with a true stick/slip type interface), it's nice to define the matrix with the dimension associated with the mesh+basis and leave all bc's in the operator. Leaving the bc's in the operator can be implemented in a manner which still retains the operators symmetry (assuming it was symmetric to begin with). This leaves the choice of what to stick on the diagonal. Simply using "1" could screw up the spectrum of the matrix and kill the iterative solver performance. A better choice would be to insert a diagonal entry closely related to the operator; e.g. something that looks like the diagonal entry of $\int 2\eta\epsilon(u) : \epsilon(v)dV$ (for the discrete stress tensor term).*

Removing Dirichlet bc's entirely for the discrete operator sounds attractive. The code will like the FE theory and you will only be solving for variables which are "unknowns" (compared with the above). However, introducing a time dependent Dirichlet domain means the matrix must be re-sized, as should its non-zero structure be re-allocated. Also, implemented multi-grid is annoying when the Dirichlet entries are removed. In fact, most of the code associated with stripping out Diriclet bc's is annoying and ugly. However, removing the bcs ensures symmetry, it ensures the discrete operator will have a nice spectrum (c.f. the above option). Also, stripping out bcs usually increases overall storage as you have one representation of the discrete vectors given to the solver which will be of size (N-n) and in your mesh you will have a repsentation of length N. "N" being the total number of dofs in your system, "n" being the number of Dirichlet constrained dofs in your system.

10.14.2 In-out flux boundary conditions for lithospheric models
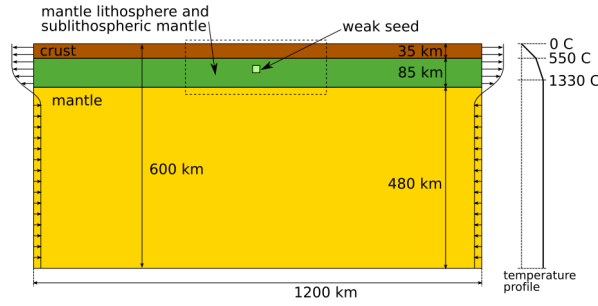
The velocity on the side is given by

$$
\begin{aligned}
u(y) &= u_{out} & y < y_1 \\
u(y) &= \frac{u_{in} - u_{out}}{y_2 - y_1}(y - y_1) + u_{out} & y_1 < y < y_2 \\
u(y) &= u_{in} & y > y_2
\end{aligned}
$$

Note that u_{in} and u_{out} can be positive or negative, but of opposite signs. The requirement for volume conservation is:

$$
\Phi = \int_0^{L_y} u(y)dy = 0
$$

Having chosen u_{in} (the velocity of the plate), one can then compute v_{out} as a function of y_1 and y_2.

$$
\Phi = \underbrace{\int_0^{y_1} u(y)dy}_{\Phi_1} + \underbrace{\int_{y_1}^{y_2} u(y)dy}_{\Phi_2} + \underbrace{\int_{y_2}^{L_y} u(y)dy}_{\Phi_3}
$$

510

with

$$\Phi_1 = u_{out}y_1$$

$$\Phi_2 = \int_{y_1}^{y_2} u(y)dy$$

$$= \int_{y_1}^{y_2} \left[\frac{u_{in} - u_{out}}{y_2 - y_1}(y - y_1) + u_{out} \right] dy$$

$$= \int_{y_1}^{y_2} \frac{u_{in} - u_{out}}{y_2 - y_1}(y - y_1)dy + \int_{y_1}^{y_2} u_{out}dy$$

$$= \frac{u_{in} - u_{out}}{y_2 - y_1} \int_{y_1}^{y_2} (y - y_1)dy + (y_2 - y_1)u_{out}$$

$$= \frac{u_{in} - u_{out}}{y_2 - y_1} \left[\frac{1}{2}y^2 - y_1 y \right]_{y_1}^{y_2} + (y_2 - y_1)u_{out}$$

$$= \frac{u_{in} - u_{out}}{y_2 - y_1}(\frac{1}{2}y_2^2 - y_1 y_2 - \frac{1}{2}y_1^2 + y_1^2) + (y_2 - y_1)u_{out}$$

$$= \frac{u_{in} - u_{out}}{y_2 - y_1}\frac{1}{2}(y_2^2 - 2y_1 y_2 + y_1^2) + (y_2 - y_1)u_{out}$$

$$= \frac{u_{in} - u_{out}}{y_2 - y_1}\frac{1}{2}(y_2 - y_1)^2 + (y_2 - y_1)u_{out}$$

$$= (u_{in} - u_{out})\frac{1}{2}(y_2 - y_1) + (y_2 - y_1)u_{out}$$

$$= \frac{1}{2}(u_{in} + u_{out})(y_2 - y_1)$$

$$\Phi_3 = (L_y - y_2)u_{in}$$

so that in the end

$$\Phi = \Phi_1 + \Phi_2 + \Phi_3$$

$$= u_{out}[y_1 + \frac{1}{2}(y_2 - y_1)] + u_{in}[\frac{1}{2}(y_2 - y_1) + (L_y - y_2)]$$

$$= u_{out}\frac{1}{2}(y_1 + y_2) + u_{in}[L_y - \frac{1}{2}(y_1 + y_2)]$$

and finally, solving for u_{out}

$$u_{out} = -u_{in}\frac{L_y - \frac{1}{2}(y_1 + y_2)}{\frac{1}{2}(y_1 + y_2)}$$

In some cases one may wish to prescribe a zero velocity below the 660 discontinuity (given by $y = y_1$) on the following figure:
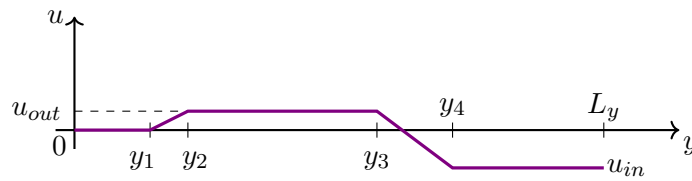
10.14.3   Periodic boundary conditions

This type of boundary conditions can be handy in some specific cases such as infinite domains. The idea is simple: when material leaves the domain through a boundary it comes back in through the opposite boundary (which of course presupposes a certain topology of the domain).

For instance, if one wants to model a gas at the molecular level and wishes to avoid interactions of the molecules with the walls of the container, such boundary conditions can be used, mimicking an infinite domain in all directions.

Let us consider the small mesh depicted hereunder:

missing picture

We wish to implement horizontal boundary conditions so that

$$u_5 = u_1 \qquad u_{10} = u_6 \qquad u_{15} = u_{11} \qquad u_{20} = u_{16}$$

One could of course rewrite these conditions as constraints and extend the Stokes matrix but this approach turns out to be not practical at all.

Instead, the method is rather simple: replace in the connectivity array the dofs on the right side (nodes 5, 10, 15, 20) by the dofs on the left side. In essence, we wrap the system upon itself in the horizontal direction so that elements 4, 8 and 12 'see' and are 'made of' the nodes 1, 6, 11 and 16. In fact, this is only necessary during the assembly. Everywhere in the loops nodes 5, 10, 15 and 20 appear one must replace them by their left pendants 1, 6, 11 and 16. This autmatically generates a matrix with lines and columns corresponding to the u_5, u_{10}, u_{15} and u_{20} being exactly zero. The Stokes matrix is the same size, the blocks are the same size and the symmetric character of the matrix is respected. However, there remains a problem. There are zeros on the diagonal of the above mentioned lines and columns. One must then place there 1 or a more appropriate value.

Another way of seeing this is as follows: let us assume we have built and assembled the Stokes matrix, and we want to impose periodic b.c. so that dof j and i are the same. The algorithm is composed of four steps:

1. add col j to col i

2. add row j to row i (including rhs)

3. zero out row j, col j

4. put average diagonal value on diagonal (j, j)

Remark. *Unfortunately the non-zero pattern of the matrix with periodic b.c. is not the same as the matrix without periodic b.c.*

10.14.4 Free-slip boundary conditions on annulus

In the context of geodynamical modelling we often wish to prescribed free-slip boundary conditions on a given boundary of the domain. If the domain is a rectangle which sides align with the Cartesian axis, then fixing $\mathbf{v}_x = 0$ or $\mathbf{v}_y = 0$ is simple and does indeed insure free-slip boundary conditions.

However the situation is much more complicated in the case of a curved boundary, such as for instance the inner and outer boundaries of an annulus or spherical shell.

If the curved boundary is a circular, the procedure is as follows:

1. identify the node on the boundary which is to be fixed.

2. compute its coordinate angle θ (and ϕ in 3D)

3. do a rotation so as to bring it back onto the x-axis (2D) or z-axis (3D)

4. apply free slip boundary condition (now easy since parallel or perpendicular to axis)

5. rotate back

This technique is implemented in STONE **??**, STONE **??** and STONE **??**.

A few remarks about rotation matrices   In a given plane, the counter-clockwise rotation matrix by and angle θ is defined by

$$\mathcal{R} = \left(\begin{array}{cc} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{array} \right)$$

The image of vector $\vec{V}$ by a rotation of angle θ is given by

$$\vec{V}' = \mathcal{R} \cdot \vec{V}$$

Coordinate transformations of second-rank tensors involve the very same matrix as vector transforms. A transformation of the stress tensor $\boldsymbol{\sigma}$, from the reference xy-coordinate system to $\boldsymbol{\sigma}'$ in a new $x'y'-$system is done as follows:

$$\boldsymbol{\sigma}' = \mathcal{R} \cdot \boldsymbol{\sigma} \cdot \mathcal{R}^T$$

[from Wikipedia] A basic rotation (also called elemental rotation) is a rotation about one of the axes of a Coordinate system. The following three basic rotation matrices rotate vectors by an angle α about the x-, y-, or z-axis, in three dimensions, using the right-hand rule which codifies their alternating signs.

$$\mathcal{R}_x(\alpha) = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{array} \right)$$

$$\mathcal{R}_y(\alpha) = \left(\begin{array}{ccc} \cos\alpha & 0 & \sin\alpha \\ 0 & 1 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{array} \right)$$

$$\mathcal{R}_z(\alpha) = \left(\begin{array}{ccc} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{array} \right)$$

In my ELEFANT code I first rotate around the z axis by and angle $-\phi$ and then around axis y by an angle $-\theta$ in the case of a spherical shell.

$$\mathcal{R}_y(-\theta) = \begin{pmatrix} \cos(-\theta) & 0 & \sin(-\theta) \\ 0 & 1 & 0 \\ -\sin(-\theta) & 0 & \cos(-\theta) \end{pmatrix} = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$$

$$\mathcal{R}_z(-\phi) = \begin{pmatrix} \cos(-\phi) & -\sin(-\phi) & 0 \\ \sin(-\phi) & \cos(-\phi) & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

These are the `Rott` and `Rotp` matrices in the routines.

📚 Relevant Literature

- Note that in some cases applying free slip boundary conditions on a curved boundary with a triangular mesh can be problematic as explained in Dione, Tibirna, and Urquiza [355] (2013).

- M.S. Engelman, R.L. Sani, and P.M. Gresho. "The implementation of normal and/or tangential boundary conditions in finite element codes for incompressible fluid flow". In: *Int. J. Num. Meth. Fluids* 2 (1982), pp. 225–238. DOI: `10.1002/fld.1650020302`

- M. Behr. "On the Application of Slip Boundary Condition on Curved Boundaries". In: *Int. J. Num. Meth. Fluids* 45 (2004), pp. 43–51. DOI: `10.1002/fld.663` in which it is stated:
  1. If the slip boundary coincides with a Cartesian coordinate plane, the implementation is trivial, with the equations corresponding to the normal component of velocity simply being dropped from the equation system. 2. If the slip boundary does not coincide with a Cartesian coordinate plane, the equations corresponding to the velocity components at the boundary are locally aligned with the normal- tangent-bi-tangent coordinate system, and the normal component of velocity is set to zero. This procedure is described by Engelman, Sani, and Gresho [395] (1982), who also advocate the use of consistent normals for proper mass conservation.

10.15 Open boundary conditions

So-called open boundary conditions have a special meaning in computational geodynamics. They usually refer to the boundary conditions on the sides of Cartesian models, usually looking at subduction or rifting processes.

In the literature boundary conditions on the vertical sidewalls are usually

- no-slip (no flow at the boundary),

- free slip (impermeable);

- open to some particular form of through-flow.

Free slip is the most commonly used boundary condition while prescribed in- and outflow or periodic boundary conditions are also common. (REF?)

Taken from Chertova, Geenen, van den Berg, and Spakman [249] (2012): "Open boundaries for which the horizontal in- and outflow are defined by a fully internally developed flow, have hardly been used [...]. Such open boundaries basically prescribe a hydrostatic pressure condition on the boundary preventing the model to collapse while horizontal in and outflow is free, in the sense that it is driven by the internal dynamics and the usual condition of incompressible flow. Among the range of boundary conditions used, open boundaries may fit best to real-mantle flow conditions surrounding subduction zones."

Two examples of the use of such boundary conditions were found in the literature: Quinteros, Sobolev, and Popov [1096] (2010) and Chertova, Geenen, van den Berg, and Spakman [249] (212).

We start again from the variational form of the momentum equation, and focus on the term containing the full stress tensor $\boldsymbol{\sigma}$. Let us look at the stress tensor gradient, multiplied by the basis function $\mathcal{N}$, integrated over the domain:

$$
\begin{aligned}
\int_V \mathcal{N} \vec{\nabla} \cdot \boldsymbol{\sigma} \; dV &= \int_V \left[\vec{\nabla} \cdot (\mathcal{N} \boldsymbol{\sigma}) - \vec{\nabla} \mathcal{N} \cdot \boldsymbol{\sigma} \right] \; dV \\
&= \int_V \vec{\nabla} \cdot (\mathcal{N} \boldsymbol{\sigma}) \; dV - \int_V \vec{\nabla} \mathcal{N} \cdot \boldsymbol{\sigma} \; dV
\end{aligned}
\tag{10.235}
$$

The right term yields the $\mathbb{K}$ and $\mathbb{G}$ matrices after discretisation, as seen in Section **??**. Turning to the left term, we then make use of the Green-Gauss divergence theorem[47] which states that for a continuously differentiable vector field $\vec{F}$:

$$
\int_V (\vec{\nabla} \cdot \vec{F}) \; dV = \int_S \vec{F} \cdot \vec{n} \; dS
$$

so that (applying it now to tensors):

$$
\int_V \vec{\nabla} \cdot (\mathcal{N} \boldsymbol{\sigma}) \; dV = \int_S \mathcal{N} \boldsymbol{\sigma} \cdot \vec{n} \; dS
$$

This right hand side term is responsible for the surface boundary conditions and cannot be neglected if one wishes to implement stress boundary conditions, such as the so-called open boundary conditions.

Note that in Liao and Gerya [832] (2017) the authors describe an iterative algorithm that allows them to control the actual force applied at the boundary by scaling the kinematical boundary conditions

---

[47]https://en.wikipedia.org/wiki/Divergence_theorem

10.15.1 Two-dimensional case - $Q_1 \times P_0$ elements

On the following figure two elements are represented, one on the left boundary, one on the right boundary:
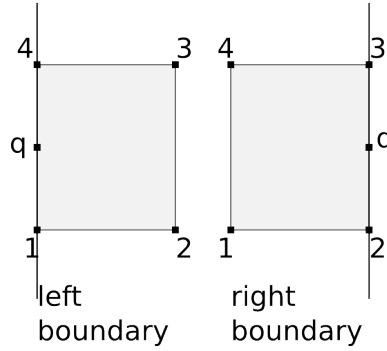
The prescribed traction on the leftt boundary is

$$\vec{t} = \boldsymbol{\sigma} \cdot \vec{n} = \begin{pmatrix} -p_{bc} & 0 \\ 0 & -p_{bc} \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_{bc} \\ 0 \end{pmatrix}$$

The integral on the side of the element is then

$$\int_\Gamma \mathcal{N}_i \vec{t} \, dS$$

for $i = 1, 2, 3, 4$, which yields the following elemental rhs vector:

$$\vec{F}_{el} = \int_{\Gamma_{14}} \begin{pmatrix} \mathcal{N}_1(x,y) t_x(x,y) \\ \mathcal{N}_1(x,y) t_y(x,y) \\ \mathcal{N}_2(x,y) t_x(x,y) \\ \mathcal{N}_2(x,y) t_y(x,y) \\ \mathcal{N}_3(x,y) t_x(x,y) \\ \mathcal{N}_3(x,y) t_y(x,y) \\ \mathcal{N}_4(x,y) t_x(x,y) \\ \mathcal{N}_4(x,y) t_y(x,y) \end{pmatrix} dS$$

It is worth noting that the integral takes place on the edge Γ_{14} so that $\mathcal{N}_2$ and $\mathcal{N}_3$ are identically zero on this edge and also $t_y = 0$ so

$$\vec{F}_{el} = \begin{pmatrix} \int_{\Gamma_{14}} \mathcal{N}_1(x,y) t_x(x,y) dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{14}} \mathcal{N}_4(x,y) t_x(x,y) dS \\ 0 \end{pmatrix}$$

If the traction (applied pressure) is constant over the element, then

$$\vec{F}_{el} = t_x \begin{pmatrix} \int_{\Gamma_{14}} \mathcal{N}_1(x,y) dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{14}} \mathcal{N}_4(x,y) dS \\ 0 \end{pmatrix} = t_x \begin{pmatrix} \int_{y_1}^{y_4} \mathcal{N}_1(x,y) dy \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{y_1}^{y_4} \mathcal{N}_4(x,y) dy \\ 0 \end{pmatrix} = \frac{t_x h_y}{2} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

516

where h_y is the height of the element along the segment.

On the right boundary, we have $\mathcal{N}_2 = 0$ and $\mathcal{N}_3 = 0$, and since $t_y = 0$ then the corresponding additional elemental right hand side vector writes:

$$\vec{F}_{el} = -\frac{t_x h_y}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

In the case where the traction is not constant over the edge, a numerical quadrature rule must be employed to integrate $\int_\Gamma \mathcal{N}_i t_x dS$.

10.15.2 Three-dimensional case - $Q_1 \times P_0$ elements

The right hand side is $ndof \times ndim = 8 \times 3 = 24$ long.
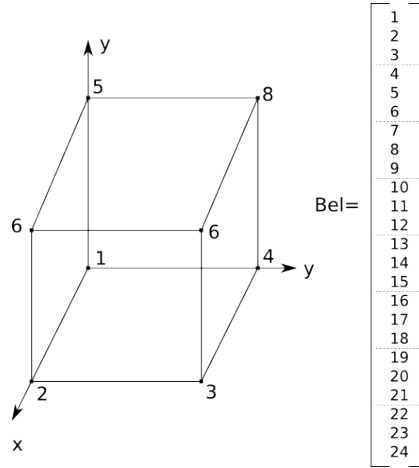
- The face $r = -1$ is made of nodes 1,4,5,8, so $\vec{n} = (-1, 0, 0)$. Since $t_y = 0$ and $t_z = 0$ and $\mathcal{N}_2 = \mathcal{N}_3 = \mathcal{N}_6 = \mathcal{N}_7$ on this face:

$$\vec{F}_{el} = \int_{\Gamma_{1458}} \begin{pmatrix} \mathcal{N}_1(x,y,z)t_x(x,y,z) \\ \mathcal{N}_1(x,y,z)t_y(x,y,z) \\ \mathcal{N}_1(x,y,z)t_z(x,y,z) \\ \mathcal{N}_2(x,y,z)t_x(x,y,z) \\ \mathcal{N}_2(x,y,z)t_y(x,y,z) \\ \mathcal{N}_2(x,y,z)t_z(x,y,z) \\ \mathcal{N}_3(x,y,z)t_x(x,y,z) \\ \mathcal{N}_3(x,y,z)t_y(x,y,z) \\ \mathcal{N}_3(x,y,z)t_z(x,y,z) \\ \mathcal{N}_4(x,y,z)t_x(x,y,z) \\ \mathcal{N}_4(x,y,z)t_y(x,y,z) \\ \mathcal{N}_4(x,y,z)t_z(x,y,z) \\ \mathcal{N}_5(x,y,z)t_x(x,y,z) \\ \mathcal{N}_5(x,y,z)t_y(x,y,z) \\ \mathcal{N}_5(x,y,z)t_z(x,y,z) \\ \mathcal{N}_6(x,y,z)t_x(x,y,z) \\ \mathcal{N}_6(x,y,z)t_y(x,y,z) \\ \mathcal{N}_6(x,y,z)t_z(x,y,z) \\ \mathcal{N}_7(x,y,z)t_x(x,y,z) \\ \mathcal{N}_7(x,y,z)t_y(x,y,z) \\ \mathcal{N}_7(x,y,z)t_z(x,y,z) \\ \mathcal{N}_8(x,y,z)t_x(x,y,z) \\ \mathcal{N}_8(x,y,z)t_y(x,y,z) \\ \mathcal{N}_8(x,y,z)t_z(x,y,z) \end{pmatrix} dS = \int_{\Gamma_{1458}} \begin{pmatrix} \mathcal{N}_1(x,y,z)t_x \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \mathcal{N}_4(x,y,z)t_x \\ 0 \\ 0 \\ \mathcal{N}_5(x,y,z)t_x \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \mathcal{N}_8(x,y,z)t_x \\ 0 \\ 0 \end{pmatrix} dS = t_x \begin{pmatrix} \int_{\Gamma_{1458}} \mathcal{N}_1(x,y,z)dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{1458}} \mathcal{N}_4(x,y,z)dS \\ 0 \\ 0 \\ \int_{\Gamma_{1458}} \mathcal{N}_5(x,y,z)dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{1458}} \mathcal{N}_8(x,y,z)dS \\ 0 \\ 0 \end{pmatrix} = \frac{h_y h_z t_x}{4} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

- The face $r = +1$ is made of nodes 2,3,6,7, so $\vec{n} = (1, 0, 0)$. so the non-zero terms are in positions $(4, 7, 16, 19)$.

- The face $s = -1$ is made of nodes 1,2,5,6, so $\vec{n} = (0, -1, 0)$. so the non-zero terms are in positions $(2, 5, 14, 17)$.

- The face $s = +1$ is made of nodes 3,4,7,8, so $\vec{n} = (0, +1, 0)$. so the non-zero terms are in positions $(8, 11, 20, 23)$.

10.15.3 Two-dimensional case - $Q_2 \times Q_1$ elements

We here too assume that we wish to prescribe a traction on the sides of a 2D domain which are aligned with the vertical axis.

constant traction

It is not fundamentally different, except that the element counts 9 nodes, so the vector is $9 \times 2 = 18$ long. The internal numbering of the nodes is as follows:

```
velocity      pressure
3---6---2     3-------2
|       |     |       |
7   8   5     |       |
|       |     |       |
0---4---1     0-------1
```

On the left boundary nodes 0,3,7 are involved while on the right boundary nodes 1,2,5 are. Assuming once again t_x constant over the edge and $t_y = 0$, we have on the left side:

$$\vec{F}_{el} = \int_{\Gamma_{073}} \begin{pmatrix} \mathcal{N}_0(x,y)t_x(x,y) \\ \mathcal{N}_0(x,y)t_y(x,y) \\ \mathcal{N}_1(x,y)t_x(x,y) \\ \mathcal{N}_1(x,y)t_y(x,y) \\ \mathcal{N}_2(x,y)t_x(x,y) \\ \mathcal{N}_2(x,y)t_y(x,y) \\ \mathcal{N}_3(x,y)t_x(x,y) \\ \mathcal{N}_3(x,y)t_y(x,y) \\ \mathcal{N}_4(x,y)t_x(x,y) \\ \mathcal{N}_4(x,y)t_y(x,y) \\ \mathcal{N}_5(x,y)t_x(x,y) \\ \mathcal{N}_5(x,y)t_y(x,y) \\ \mathcal{N}_6(x,y)t_x(x,y) \\ \mathcal{N}_6(x,y)t_y(x,y) \\ \mathcal{N}_7(x,y)t_x(x,y) \\ \mathcal{N}_7(x,y)t_y(x,y) \\ \mathcal{N}_8(x,y)t_x(x,y) \\ \mathcal{N}_8(x,y)t_y(x,y) \end{pmatrix} dS = t_x \begin{pmatrix} \int_{\Gamma_{073}} \mathcal{N}_0(x,y)dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{073}} \mathcal{N}_3(x,y)dS \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{\Gamma_{073}} \mathcal{N}_7(x,y)dS \\ 0 \\ 0 \\ 0 \end{pmatrix} = t_x \frac{h_y}{2} \begin{pmatrix} \int_{-1}^{+1} \mathcal{N}_0(r=-1,s)ds \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{-1}^{+1} \mathcal{N}_3(r=-1,s)ds \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \int_{-1}^{+1} \mathcal{N}_7(r=-1,s)ds \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We then compute

$$\int_{-1}^{+1} \mathcal{N}_0(r=-1,s)ds = \int_{-1}^{+1} \frac{1}{2}s(s-1)ds = \frac{1}{3} \tag{10.236}$$

$$\int_{-1}^{+1} \mathcal{N}_3(r=-1,s)ds = \int_{-1}^{+1} \frac{1}{2}s(s+1)ds = \frac{1}{3} \tag{10.237}$$

$$\int_{-1}^{+1} \mathcal{N}_7(r=-1,s)ds = \int_{-1}^{+1} (1-s^2)ds = \frac{4}{3} \tag{10.238}$$

Note that the sum of the three terms is 2, as expected: on the edge we have $\mathcal{N}_0 + \mathcal{N}_3 + \mathcal{N}_7 = 1$ so that the integral of the sum over the interval [-1,1] yields 2. Finally

$$
\vec{F}_{el} = \frac{t_x h_y}{6}
\begin{pmatrix}
1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 0
\end{pmatrix}
$$

This is implemented in STONE 61, 64, 146, 148.

On the right boundary, we need to compute (careful with sign when implementing!)

$$
\vec{F}_{el} = \int_{\Gamma_{152}}
\begin{pmatrix}
\mathcal{N}_0(x,y)t_x(x,y) \\
\mathcal{N}_0(x,y)t_y(x,y) \\
\mathcal{N}_1(x,y)t_x(x,y) \\
\mathcal{N}_1(x,y)t_y(x,y) \\
\mathcal{N}_2(x,y)t_x(x,y) \\
\mathcal{N}_2(x,y)t_y(x,y) \\
\mathcal{N}_3(x,y)t_x(x,y) \\
\mathcal{N}_3(x,y)t_y(x,y) \\
\mathcal{N}_4(x,y)t_x(x,y) \\
\mathcal{N}_4(x,y)t_y(x,y) \\
\mathcal{N}_5(x,y)t_x(x,y) \\
\mathcal{N}_5(x,y)t_y(x,y) \\
\mathcal{N}_6(x,y)t_x(x,y) \\
\mathcal{N}_6(x,y)t_y(x,y) \\
\mathcal{N}_7(x,y)t_x(x,y) \\
\mathcal{N}_7(x,y)t_y(x,y) \\
\mathcal{N}_8(x,y)t_x(x,y) \\
\mathcal{N}_8(x,y)t_y(x,y)
\end{pmatrix}
dS = t_x
\begin{pmatrix}
0 \\
0 \\
\int_{\Gamma_{125}} \mathcal{N}_1(x,y)dS \\
0 \\
\int_{\Gamma_{125}} \mathcal{N}_2(x,y)dS \\
0 \\
0 \\
0 \\
0 \\
0 \\
\int_{\Gamma_{125}} \mathcal{N}_5(x,y)dS \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}
= t_x \frac{h_y}{2}
\begin{pmatrix}
0 \\
0 \\
\int_{-1}^{+1} \mathcal{N}_1(-1,s)ds \\
0 \\
\int_{-1}^{+1} \mathcal{N}_2(-1,s)ds \\
0 \\
0 \\
0 \\
0 \\
0 \\
\int_{-1}^{+1} \mathcal{N}_5(-1,s)ds \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}
= \frac{t_x h_y}{6}
\begin{pmatrix}
0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{pmatrix}
$$

linear traction

Let us now turn to the case where the traction we wish to apply on the boundary is not piecewise constant but linear. We set $t_x(y) = ay + b$, so that on the right side (nodes, 1,2,5), we have to compute

$$\int_{\Gamma_{125}} \mathcal{N}_1(x,y) t_x(y) dS = \int_{\Gamma_{125}} \mathcal{N}_1(x,y)(ay+b) dy$$

$$= \frac{h_y}{2} \int_{-1}^{1} \mathcal{N}_1(r=1,s)[ay(s)+b] ds$$

We have

$$s(y) = \frac{2}{h_y}(y - y_1) - 1 \qquad \text{or} \qquad y(s) = \frac{h_y}{2}(s+1) + y_1$$

and

$$\mathcal{N}_1(r,s) = \frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \qquad \Rightarrow \qquad \mathcal{N}_1(r=1,s) = \frac{1}{2}1(1+1)\frac{1}{2}s(s-1) = \frac{1}{2}s(s-1)$$

Then

$$\int_{\Gamma_{125}} \mathcal{N}_1(x,y) t_x(y) dS = \frac{h_y}{2} \int_{-1}^{1} \mathcal{N}_1(r=1,s) \left[a\left(\frac{h_y}{2}(s+1)+y_1\right)+b \right] ds$$

$$= \frac{h_y}{2} \int_{-1}^{1} \frac{1}{2}s(s-1) \left[a\left(\frac{h_y}{2}(s+1)+y_1\right)+b \right] ds$$

$$= \frac{h_y}{4} \int_{-1}^{1} s(s-1) \left[\frac{ah_y}{2}(s+1)+(ay_1+b) \right] ds$$

$$= \frac{h_y}{4} \left[\frac{ah_y}{2} \int_{-1}^{1} s(s-1)(s+1) ds + (ay_1+b) \int_{-1}^{1} s(s-1) ds \right]$$

$$= \frac{h_y}{4}\frac{ah_y}{2} \underbrace{\int_{-1}^{1} s(s^2-1) ds}_{=0} + \frac{h_y}{4}(ay_1+b) \underbrace{\int_{-1}^{1} s(s-1) ds}_{=2/3}$$

$$= \frac{h_y}{6}(ay_1+b) \qquad\qquad (10.239)$$

Let us now turn to $\mathcal{N}_2$:

$$\mathcal{N}_2(r,s) = \frac{1}{2}r(r+1)\frac{1}{2}s(s+1) \qquad \Rightarrow \qquad \mathcal{N}_2(r=1,s) = \frac{1}{2}s(s+1)$$

Then

$$\int_{\Gamma_{125}} \mathcal{N}_2(x,y) t_x(y) dS = \frac{h_y}{2} \int_{-1}^{1} \mathcal{N}_2(r=1,s) \left[a\left(\frac{h_y}{2}(s+1)+y_1\right)+b \right] ds$$

$$= \frac{h_y}{2} \int_{-1}^{1} \frac{1}{2}s(s+1) \left[a\left(\frac{h_y}{2}(s+1)+y_1\right)+b \right] ds$$

$$= \frac{h_y}{4} \int_{-1}^{1} s(s+1) \left[\frac{ah_y}{2}(s+1)+(ay_1+b) \right] ds$$

$$= \frac{h_y}{4} \left[\frac{ah_y}{2} \int_{-1}^{1} s(s+1)(s+1) ds + (ay_1+b) \int_{-1}^{1} s(s+1) ds \right]$$

$$= \frac{h_y}{4}\frac{ah_y}{2} \underbrace{\int_{-1}^{1} s(s+1)^2 ds}_{=4/3} + \frac{h_y}{4}(ay_1+b) \underbrace{\int_{-1}^{1} s(s+1) ds}_{=2/3}$$

$$= \frac{h_y}{6}(ah_y+ay_1+b) \qquad\qquad (10.240)$$

And finally let us turn to $\mathcal{N}_5$:

$$\mathcal{N}_5(r,s) = \frac{1}{2}r(r+1)(1-s^2) \qquad \Rightarrow \qquad \mathcal{N}_5(r=1,s) = (1-s^2)$$

then

$$
\begin{aligned}
\int_{\Gamma_{125}} \mathcal{N}_5(x,y)t_x(y)dS &= \frac{h_y}{2} \int_{-1}^{1} \mathcal{N}_2(r=1,s) \left[a\left(\frac{h_y}{2}(s+1) + y_1 \right) + b \right] ds \\
&= \frac{h_y}{2} \int_{-1}^{1} (1-s^2) \left[a\left(\frac{h_y}{2}(s+1) + y_1 \right) + b \right] ds \\
&= \frac{h_y}{2} \int_{-1}^{1} (1-s^2) \left[\frac{ah_y}{2}(s+1) + (ay_1 + b) \right] ds \\
&= \frac{h_y}{2} \left[\frac{ah_y}{2} \int_{-1}^{1} (1-s^2)(s+1)ds + (ay_1 + b) \int_{-1}^{1} (1-s^2)ds \right] \\
&= \frac{h_y}{2} \frac{ah_y}{2} \underbrace{\int_{-1}^{1} (1-s^2)(1+s)ds}_{=4/3} + \frac{h_y}{2}(ay_1+b) \underbrace{\int_{-1}^{1} (1-s^2)ds}_{=4/3} \\
&= \frac{h_y}{6} \left(2ah_y + 4ay_1 + 4b \right)
\end{aligned}
\tag{10.241}
$$

Note that by setting $a=0$ and $b=t_x$ we recover the expressions above for a piecewise constant value.

If we know p_1 and p_2 (say, for example that the lithostatic pressure has been computed on these nodes and we wish to prescribe it on the side) then

$$t_y = ay + b = \underbrace{\frac{p_2 - p_1}{y_2 - y_1}}_{=a} y + \underbrace{p_1 - \frac{p_2 - p_1}{y_2 - y_1}y_1}_{=b}$$

On the left side (nodes 0,7,3), we have to compute

$$\int_{073} \mathcal{N}_0(x,y)(ay+b)dS = \frac{h_y}{2} \int_{-1}^{+1} \mathcal{N}_0(r=-1,s)[ay(s)+b]ds \tag{10.242}$$

$$\int_{073} \mathcal{N}_3(x,y)(ay+b)dS = \frac{h_y}{2} \int_{-1}^{+1} \mathcal{N}_3(r=-1,s)[ay(s)+b]ds \tag{10.243}$$

$$\int_{073} \mathcal{N}_7(x,y)(ay+b)dS = \frac{h_y}{2} \int_{-1}^{+1} \mathcal{N}_7(r=-1,s)[ay(s)+b]ds \tag{10.244}$$

with

$$\mathcal{N}_0(r,s) = \frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \to \mathcal{N}_0(-1,s) \qquad = \frac{1}{2}s(s-1)$$

$$\mathcal{N}_3(r,s) = \frac{1}{2}r(r-1)\frac{1}{2}(1-s^2) \to \mathcal{N}_3(-1,s) \qquad = (1-s^2)$$

$$\mathcal{N}_7(r,s) = \frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \to \mathcal{N}_7(-1,s) \qquad = \frac{1}{2}s(s+1)$$

If we know p_0 and p_3 then

$$t_x = ay + b = \underbrace{\frac{p_3 - p_0}{y_3 - y_0}}_{} = ay + \underbrace{p_0 - \frac{p_3 - p_0}{y_3 - y_0}y_0}_{=b}$$

10.15.4 Two-dimensional case - Linear triangle elements

Let us assume we want to apply a stress on the face 13 of the following element:

```
1------------3
 \          /
  \        /
   \      /
    \    /
     \  /
      \/
      2
```
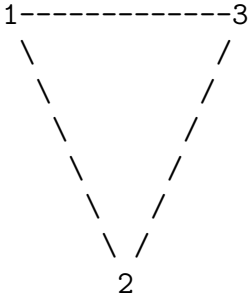
The integral on the side of the element is $\int_\Gamma \mathcal{N}_i \vec{t} \, dS$ for $i = 1, 2, 3$, which yields the following elemental rhs vector:

$$\vec{F}_{el} = \int_{\Gamma_{13}} \begin{pmatrix} \mathcal{N}_1(x,y)t_x(x,y) \\ \mathcal{N}_1(x,y)t_y(x,y) \\ \mathcal{N}_2(x,y)t_x(x,y) \\ \mathcal{N}_2(x,y)t_y(x,y) \\ \mathcal{N}_3(x,y)t_x(x,y) \\ \mathcal{N}_3(x,y)t_y(x,y) \end{pmatrix} dS = \int_{\Gamma_{13}} \begin{pmatrix} 0 \\ \mathcal{N}_1(x,y)t_y(x,y) \\ 0 \\ 0 \\ 0 \\ \mathcal{N}_3(x,y)t_y(x,y) \end{pmatrix} dS$$

since $t_x = 0$ and there function $\mathcal{N}_2$ will be zero on the edge.

We also arbitrarily set $y_1 = y_3 = 0$. We have seen in Section ?? that the basis functions (expressed as a function of the real coordinates x, y) for a linear triangle are given by:

$$\mathcal{N}_1(x,y) = \frac{1}{D}[(x_2 y_3 - x_3 y_2) + (y_2 - y_3)x + (x_3 - x_2)y]$$

$$\mathcal{N}_2(x,y) = \frac{1}{D}[(x_3 y_1 - x_1 y_3) + (y_3 - y_1)x + (x_1 - x_3)y]$$

$$\mathcal{N}_3(x,y) = \frac{1}{D}[(x_1 y_2 - x_2 y_1) + (y_1 - y_2)x + (x_2 - x_1)y]$$

with

$$D = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = \begin{vmatrix} 1 & x_1 & 0 \\ 1 & x_2 & y_2 \\ 1 & x_3 & 0 \end{vmatrix} = -x_3 y_2 + x_1 y_2 = y_2(x_1 - x_3)$$

$$
\begin{aligned}
\int_{x_1}^{x_3} \mathcal{N}_1(x, y = 0)dx &= \frac{1}{D} \int_{x_1}^{x_3} [(x_2y_3 - x_3y_2) + (y_2 - y_3)x] \\
&= \frac{1}{D} \int_{x_1}^{x_3} [-x_3y_2 + y_2x] \qquad \text{since} y_1 = y_3 = 0 \\
&= \frac{y_2}{D} \int_{x_1}^{x_3} (-x_3 + x)dx \\
&= \frac{y_2}{y_2(x_1 - x_3)}[-x_3x + \frac{1}{2}x^2]_{x_1}^{x_3} \\
&= \frac{1}{x_1 - x_3}[-x_3(x_3 - x_1) + \frac{1}{2}(x_3^2 - x_1^2)] \\
&= \frac{1}{x_1 - x_3}[x_3(x_1 - x_3) + \frac{1}{2}(x_3 - x_1)(x_3 + x_1)] \\
&= x_3 - \frac{1}{2}(x_3 + x_1) \\
&= \frac{1}{2}(x_3 - x_1) \qquad\qquad\qquad (10.245) \\
\int_{x_1}^{x_3} \mathcal{N}_3(x, y = 0)dx &= \frac{1}{D} \int_{x_1}^{x_3} [(x_1y_2 - x_2y_1) + (y_1 - y_2)x]dx \\
&= \frac{1}{D} \int_{x_1}^{x_3} [x_1y_2 - y_2x]dx \\
&= \frac{y_2}{D} \int_{x_1}^{x_3} [x_1 - x]dx \\
&= \frac{y_2}{y_2(x_1 - x_3)}[x_1x - \frac{1}{2}x^2]_{x_1}^{x_3} \\
&= \frac{1}{x_1 - x_3}[x_1(x_3 - x_1) - \frac{1}{2}(x_3^2 - x_1^2)] \\
&= -x_1 + \frac{1}{2}(x_3 + x_1) \\
&= \frac{1}{2}(x_3 - x_1) \qquad\qquad\qquad (10.246)
\end{aligned}
$$

Finally

$$
\vec{F}_{el} = \frac{ht_y}{2} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}
$$

10.16 About nullspaces

10.16.1 Pressure normalisation, nullspace

Basic idea and naive implementation

When Dirichlet boundary conditions are imposed everywhere on the boundary, pressure is only present by its gradient in the equations. It is thus determined up to an arbitrary constant (one speaks then of a nullspace of size 1). In such a case, one commonly impose the average of the pressure over the whole domain or on a subset of the boundary to have a zero average, i.e.

$$\int_\Omega p \, dV = 0 \tag{10.247}$$

Let us assume for example/simplicity that we are using $Q_1 \times P_0$ elements. The pressure is constant inside each element so the integral above becomes:

$$\int_\Omega p \, dV = \sum_e \int_{\Omega_e} p \, dV = \sum_e p_e \int_{\Omega_e} dV = \sum_e p_e V_e = 0 \tag{10.248}$$

where the sum runs over all elements e of volume V_e. This can be rewritten

$$\vec{L} \cdot \vec{\mathcal{P}} = 0$$

and it is a constraint on the pressure solution which couples *all* pressure dofs. We can associate to it a Lagrange multiplier λ so that we must solve the modified Stokes system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} & 0 \\ \mathbb{G}^T & 0 & \vec{L}^T \\ 0 & \vec{L} & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{\mathcal{V}} \\ \vec{\mathcal{P}} \\ \lambda \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{h} \\ 0 \end{pmatrix}$$

When higher order spaces are used for pressure (continuous or discontinuous) one must then carry out the above integration numerically by means of (usually) a Gauss-Legendre quadrature.

Although valid, this approach has one main disadvantage: it makes the Stokes matrix larger (although marginally so – only one row and column are added), but more importantly it prevents the use of some of the solving strategies of Section 10.13.

Remark: Iterative solvers are able to handle the indetermination for the pressure [1188, p. 57].

Implementation – the real deal

Here is what Bochev and Gunzburger [110, Section 7.6.4] have to say about this: "[...] practical implementations cheat by substituting enforcement of the true zero mean constraint by using procedures collectively known as setting the pressure datum. These procedures essentially amount to removing one degree of freedom from the pressure space. Setting the pressure datum can be accomplished in many different ways, ranging from specifying a pressure value at a grid node to more complicated approaches in which a boundary traction is specified at a single node in lieu of the velocity condition; see [16, 24, 191] and the references cited therein for more details. Not surprisingly, in practice, the simplest approach of fixing the pressure value at a node also happens to be the most widely used in practice."

The idea is actually quite simple and requires two steps:

1. remove the null space by prescribing the pressure at one location and solve the system;

2. post-process the pressure so as to arrive at a pressure field which fulfils the required normalisation (surface, volume, ...)

The reason why it works is as follows: a constant pressure value lies in the null space, so that one can add or delete any value to the pressure field without consequence. As such I can choose said constant such that the pressure at a given node/element is zero. All other computed pressures are then relative to that one. The post-processing step will redistribute a constant value to all pressures (it will shift them up or down) so that the normalising condition is respected.

🛢 Relevant Literature

In `https://scicomp.stackexchange.com/questions/27645/pressure-boundary-condition-in-lid-`
we find: *Zero mean pressure space is used for convenience when one is interested in FEA theory (basically, we cannot enforce $p(x_0) = p_0$ for $p \in L^2$ since it does not make sense); from the computational point of view, it is easier to fix one of the pressure DOFs (although you can subtract mean value at the post–processing step if you want to). When you are working w/ polynomial spaces—and this is exactly what you do in FEM -it is perfectly fine to enforce $p(x_0) = p_0$. Handle this constraint like you usually handle Dirichlet BCs (e.g., via modifying your matrix). It is also fine to ignore this constraint in some cases (e.g., Krylov solvers can do fine with this).*

`https://scicomp.stackexchange.com/questions/25134/mixed-finite-element-method-for-the-s`

10.16.2 Removing rotational nullspace

When free slip boundary conditions are prescribed in an annulus or hollow sphere geometry there exists a rotational nullspace, or in other words there exists a tangential velocity field ('pure rotation') which, if added or subtracted to the solution, generates a solution which is still the solution of the PDEs.

As in the pressure normalisation case (see section 10.16.1), the solution is simple:

1. fix the tangential velocity at *one* node on a boundary, and solve the system (the nullspace has been removed)[48]

2. post-process the solution to have the velocity field fulfill the required conditions, i.e. either a zero net angular momentum or a zero net angular velocity of the domain.

Remark. *In* ASPECT *this is available under the option "Remove nullspace = angular momentum" and "Remove nullspace = net rotation". The "angular momentum" option removes a rotation such that the net angular momentum is zero. The "net rotation" option removes the net rotation of the domain.*

Angular momentum approach In physics, velocity is not a conserved quantity, but the momentum is. n order to remove the angular momentum, we search for a rotation vector $\vec{\omega}$ such that

$$\int_\Omega \rho[\vec{r} \times (\vec{v} - \vec{\omega} \times \vec{r})] \, dV = \vec{0} \tag{10.249}$$

The angular momentum of a rigid body can be obtained from the sum of the angular momentums

---

[48]`https://scicomp.stackexchange.com/questions/3531/how-to-remove-rigid-body-motions-in-linear-elasticity`

of the particles forming the body[49]:

$$\vec{H} = \sum_i \vec{L}_i \tag{10.250}$$

$$= \sum_i \vec{r}_i \times m_i \vec{v}_i \tag{10.251}$$

$$= \sum_i \vec{r}_i \times m_i (\vec{\omega}_i \times \vec{r}_i) \tag{10.252}$$

$$= \sum_i m_i \begin{pmatrix} \sum_i m_i(y_i^2 + z_i^2) & -\sum_i m_i x_i y_i & -\sum_i m_i x_i z_i \\ -\sum_i m_i x_i y_i & \sum_i m_i(x_i^2 + z_i^2) & -\sum_i m_i y_i z_i \\ -\sum_i m_i x_i z_i & -\sum_i m_i y_i z_i & \sum_i m_i(x_i^2 + y_i^2) \end{pmatrix} \cdot \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} \tag{10.253}$$

In the continuum limit, we have:

$$\vec{H} = \int_\Omega \rho(\vec{r})\, \vec{r} \times \vec{v}\, dV \tag{10.254}$$

and the 3×3 moment of inertia tensor $\boldsymbol{I}$ (also called inertia tensor) is given by[50]

$$\boldsymbol{I} = \int_\Omega \rho(\vec{r})[\vec{r} \cdot \vec{r}\, \boldsymbol{1} - \vec{r} \times \vec{r}]dV \tag{10.255}$$

so that the above equation writes: $\vec{H} = \boldsymbol{I} \cdot \vec{\omega}$ and then $\vec{\omega} = \boldsymbol{I}^{-1} \cdot \vec{H}$.

Ultimately, at each velocity node a rotation about the rotation vector $\vec{\omega}$ is then subtracted from the velocity solution [1521, eq. 26]:

$$\vec{v}_{new} = \vec{v}_{old} - \vec{\omega} \times \vec{r} \tag{10.256}$$

For the special case of a solid sphere of constant density and radius R the tensor $\boldsymbol{I}$ becomes diagonal and we have

$$I = \frac{2}{5}mR^2$$

where m is the mass of the sphere.

The case of a hollow sphere is explained in Section 2.4.1 of Zhong *et al.* (2008) [1521].

Three dimensions

The angular momentum vector is given by:

$$\vec{H} = \int_\Omega \rho(\vec{r}) \begin{pmatrix} yw - zv \\ zu - xw \\ xv - yu \end{pmatrix} d\vec{r} = \begin{pmatrix} \int_\Omega \rho(\vec{r})(yw - zv)d\vec{r} \\ \int_\Omega \rho(\vec{r})(zu - xw)d\vec{r} \\ \int_\Omega \rho(\vec{r})(xv - yu)d\vec{r} \end{pmatrix} = \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} \tag{10.257}$$

---

[49]http://www.kwon3d.com/theory/moi/iten.html
[50]https://en.wikipedia.org/wiki/Moment_of_inertia

while the inertia tensor for a continuous body is given by

$$\mathbf{I} = \int_\Omega \rho(\vec{r})[\vec{r}\cdot\vec{r}\,\mathbf{1} - \vec{r}\times\vec{r}]d\vec{r} \tag{10.258}$$

$$= \int_\Omega \rho(\vec{r})\left[\begin{pmatrix} x^2+y^2+z^2 & 0 & 0 \\ 0 & x^2+y^2+z^2 & 0 \\ 0 & 0 & x^2+y^2+z^2 \end{pmatrix} - \begin{pmatrix} xx & xy & xz \\ yx & yy & yz \\ zx & zy & zz \end{pmatrix}\right]d\vec{r} \tag{10.259}$$

$$= \int_\Omega \rho(\vec{r})\begin{pmatrix} y^2+z^2 & -xy & -xz \\ -yx & x^2+z^2 & -yz \\ -zx & -zy & x^2+y^2 \end{pmatrix}d\vec{r} \tag{10.260}$$

$$= \begin{pmatrix} \int_\Omega \rho(\vec{r})(y^2+z^2)d\vec{r} & -\int_\Omega \rho(\vec{r})xyd\vec{r} & -\int_\Omega \rho(\vec{r})xzd\vec{r} \\ -\int_\Omega \rho(\vec{r})yxd\vec{r} & \int_\Omega \rho(\vec{r})(x^2+z^2)d\vec{r} & -\int_\Omega \rho(\vec{r})yzd\vec{r} \\ -\int_\Omega \rho(\vec{r})zxd\vec{r} & -\int_\Omega \rho(\vec{r})zyd\vec{r} & \int_\Omega \rho(\vec{r})(x^2+y^2)d\vec{r} \end{pmatrix} \tag{10.261}$$

$$= \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \tag{10.262}$$

Two dimensions

In two dimensions, flow is taking place in the (x,y) plane. This means that $\vec{r} = (x,y,0)$ and $\vec{v} = (u,v,0)$ are coplanar, and therefore that $\vec{\omega}$ is perpendicular to the plane. We have then

$$\vec{H} = \int_\Omega \rho(\vec{r})\begin{pmatrix} 0 \\ 0 \\ xv-yu \end{pmatrix}d\vec{r} = \begin{pmatrix} 0 \\ 0 \\ \int_\Omega \rho(\vec{r})(xv-yu)d\vec{r} \end{pmatrix} \tag{10.263}$$

and

$$\mathbf{I} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} = \begin{pmatrix} I_{xx} & I_{xy} & 0 \\ I_{yx} & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{pmatrix} \tag{10.264}$$

since $I_{xz} = I_{yz} = 0$ as $z = 0$, and with $I_{xx} = \int_\Omega \rho(\vec{r})y^2 d\vec{r}$ and $I_{yy} = \int_\Omega \rho(\vec{r})x^2 d\vec{r}$. The solution to $\mathbf{I}\cdot\vec{\omega} = \vec{H}$ can be easily obtained (see Appendix 3.1.5):

$$\omega_x = \frac{1}{det(\mathbf{I})}\begin{vmatrix} 0 & I_{xy} & 0 \\ 0 & I_{yy} & 0 \\ H_3 & 0 & I_{zz} \end{vmatrix} = 0 \tag{10.265}$$

$$\omega_y = \frac{1}{det(\mathbf{I})}\begin{vmatrix} I_{xx} & 0 & 0 \\ I_{yx} & 0 & 0 \\ 0 & H_z & I_{zz} \end{vmatrix} = 0 \tag{10.266}$$

$$\omega_z = \frac{1}{det(\mathbf{I})}\begin{vmatrix} I_{xx} & I_{xy} & 0 \\ I_{yx} & I_{yy} & 0 \\ 0 & 0 & H_z \end{vmatrix} \tag{10.267}$$

$$= \frac{1}{det(\mathbf{I})}\left(I_{xx}I_{yy}H_z - I_{yx}I_{xy}H_z\right) \tag{10.268}$$

$$= \frac{1}{det(\mathbf{I})}\left(I_{xx}I_{yy} - I_{yx}I_{xy}\right)H_z \tag{10.269}$$

with $det(\mathbf{I}) = I_{xx}I_{yy}I_{zz} - I_{yx}I_{xy}I_{zz} = (I_{xx}I_{yy} - I_{yx}I_{xy})I_{zz}$ and then

$$\omega_z = \frac{(I_{xx}I_{yy} - I_{yx}I_{xy})H_z}{(I_{xx}I_{yy} - I_{yx}I_{xy})I_{zz}} = \frac{H_z}{I_{zz}} = \frac{\int_\Omega \rho(\vec{r})(xv - yu)d\vec{r}}{\int_\Omega \rho(\vec{r})(x^2 + y^2)d\vec{r}}$$

Concretely, this means that in 2D one does not need to solve the system $\mathbf{I} \cdot \vec{\omega} = \vec{H}$ since only ω_z is not zero.

Then, since $\vec{r} = (x, y, z)$ and $\vec{\omega} = (0, 0, \omega_z)$:

$$\vec{v}_{new}(\vec{r}) = \vec{v}_{old} - \vec{\omega} \times \vec{r} = \begin{pmatrix} u_{old} - (-\omega_z y) \\ v_{old} - (\omega_z x) \\ 0 \end{pmatrix} \tag{10.270}$$

Chapter 11

Mapping & Jacobians

The name *isoparametric* derives from the fact that the same ('iso') functions are used as basis functions and for the mapping to the reference element.

More generally, if n_e denotes the number of nodes of an element and n_g denotes the number of nodes describing the geometry of the element, then the element is termed *subparametric* when $n_g < n_e$ and *superparametric* when $n_g > n_e$.

11.1 General case

What follows is written for the 2d case but extending it to 3d is trivial.

Any variable defined on the element can be approximated using the basis functions:

$$f^h(r,s) = \sum_i \mathcal{N}_i(r,s) f_i. \tag{11.1}$$

If we treat the coordinate variables x and y themselves as functions of (r,s), then the basis functions can be used to construct the mapping:

$$x(r,s) = \sum_{i=1}^m \mathcal{N}_i(r,s) x_i \tag{11.2}$$

$$y(r,s) = \sum_{i=1}^m \mathcal{N}_i(r,s) y_i \tag{11.3}$$

where m is the number of nodes. This is a relationship between the reduced coordinates r, s and the 'real' coordinates x, y.

Let us compute the space derivatives of these quantities:

$$\frac{\partial x}{\partial r} = \sum_i^m \frac{\partial \mathcal{N}_i}{\partial r} x_i \tag{11.4}$$

$$\frac{\partial x}{\partial s} = \sum_i^m \frac{\partial \mathcal{N}_i}{\partial s} x_i \tag{11.5}$$

$$\frac{\partial y}{\partial r} = \sum_i^m \frac{\partial \mathcal{N}_i}{\partial r} y_i \tag{11.6}$$

$$\frac{\partial y}{\partial s} = \sum_i^m \frac{\partial \mathcal{N}_i}{\partial s} y_i \tag{11.7}$$

We also have

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} \tag{11.8}$$

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial s} \tag{11.9}$$

or in matrix form:

$$\begin{pmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial s} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \end{pmatrix}}_{\boldsymbol{J}} \cdot \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

where $\boldsymbol{J}$ is called the Jacobian of the transformation. By inverting the Jacobian matrix, the desired derivatives with respect to x and y can be obtained:

$$\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \boldsymbol{J}^{-1} \cdot \begin{pmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial s} \end{pmatrix}$$

The inverse of the Jacobian matrix can be simply obtained in 2d (Cramer's rule for 2×2 matrices[1]):

$$\boldsymbol{J}^{-1} = \frac{1}{|\boldsymbol{J}|} \begin{pmatrix} \frac{\partial y}{\partial s} & -\frac{\partial y}{\partial r} \\ -\frac{\partial x}{\partial s} & \frac{\partial x}{\partial r} \end{pmatrix}$$

The presence of the determinant in the denominator implies that it cannot be zero anywhere, or in other words: the mapping is not valid if $|\boldsymbol{J}|$ is zero anywhere over the element.

Remark. *Problems also arise when the Jacobian matrix is nearly singular due to round-off errors. To avoid problems linked to badly shaped elements, it is recommended that the inside angles of an element are larger than $15°$ and less than $165°$.*

Since $x(r, s)$ and $y(r, s)$, we can also write:

$$dx = \frac{\partial x}{\partial r}dr + \frac{\partial x}{\partial s}ds$$

$$dy = \frac{\partial y}{\partial r}dr + \frac{\partial y}{\partial s}ds$$

and we see that the Jocobian matrix also appears so that

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \boldsymbol{J} \cdot \begin{pmatrix} dr \\ ds \end{pmatrix} \tag{11.10}$$

This means that integrating over the 'real' element in (x, y) space can be simply done by integrating of the reference element in the (r, s) space[2]. This is the cornerstone of most of the implementation of the Finite Element Method, the second integral being carried out by means of the Gauss-Legendre quadrature.

$$\iint_{\Omega_e} \dots dxdy = \int_{-1}^{+1} \int_{-1}^{+1} \dots |\boldsymbol{J}| \, drds \tag{11.11}$$

make link with LAVA chapter at some point

---

[1] https://en.wikipedia.org/wiki/Cramers_rule
[2] https://en.wikibooks.org/wiki/Calculus/Change_of_variables

11.2 A quick word about implementation

In two dimensions

```python
jcb = np.zeros((2, 2),dtype=np.float64)
for k in range(0,m):
    jcb[0,0]+=dNdr[k]*x[icon[k,iel]]
    jcb[0,1]+=dNdr[k]*y[icon[k,iel]]
    jcb[1,0]+=dNds[k]*x[icon[k,iel]]
    jcb[1,1]+=dNds[k]*y[icon[k,iel]]
```

This implementation is interesting because it explicitely implements the loop $\sum_i$ over nodes of Eqs. (11.4),(11.5),(11.6),(11.7) as a for loop. It is often used in (older) stones and progressively beeing replaced by the following implementation that is more compact and avoids the writing of explicit loops (which are notoriously slow in Python):

```python
jcb[0,0]=dNdr.dot(x[icon[:,iel]])
jcb[0,1]=dNdr.dot(y[icon[:,iel]])
jcb[1,0]=dNds.dot(x[icon[:,iel]])
jcb[1,1]=dNds.dot(y[icon[:,iel]])
```

11.3 Linear mapping on a triangle

$$
\begin{aligned}
x(r,s) &= \sum_{i=1}^{3} \mathcal{N}_i(r,s)x_i \\
&= \mathcal{N}_1(r,s)x_1 + \mathcal{N}_2(r,s)x_2 + \mathcal{N}_3(r,s)x_3 \\
&= (1-r-s)x_1 + (r)x_2 + (s)x_3 \\
&= x_1 + (x_2-x_1)r + (x_3-x_1)s \\
&= a_x + b_x r + c_x s \\
y &= \sum_{i=1}^{3} \mathcal{N}_i(r,s)y_i \\
&= \mathcal{N}_1(r,s)y_1 + \mathcal{N}_2(r,s)y_2 + \mathcal{N}_3(r,s)y_3 \\
&= (1-r-s)y_1 + (r)y_2 + (s)y_3 \\
&= y_1 + (y_2-y_1)r + (y_3-y_1)s \\
&= a_y + b_y r + c_y s
\end{aligned}
$$

Let us compute the space derivatives of these quantities:

$$
\begin{aligned}
\frac{\partial x}{\partial r} &= x_2 - x_1 = b_x \\
\frac{\partial x}{\partial s} &= x_3 - x_1 = c_x \\
\frac{\partial y}{\partial r} &= y_2 - y_1 = b_y \\
\frac{\partial y}{\partial s} &= y_3 - y_1 = c_y
\end{aligned}
$$

The jacobian matrix is then given by

$$
\boldsymbol{J} = \begin{pmatrix} x_2-x_1 & y_2-y_1 \\ x_3-x_1 & y_3-y_1 \end{pmatrix} = \begin{pmatrix} b_x & b_y \\ c_x & c_y \end{pmatrix}
$$

and its inverse

$$\boldsymbol{J}^{-1} = \frac{1}{b_x c_y - c_x b_y} \begin{pmatrix} c_y & -b_y \\ -c_x & b_x \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} c_y & -b_y \\ -c_x & b_x \end{pmatrix}$$

where $A = (b_x c_y - c_x b_y)/2$ is actually the area of the triangle!

The Cartesian basis function derivatives are then

$$\begin{pmatrix} \frac{\partial \mathcal{N}_1}{\partial x} \\ \frac{\partial \mathcal{N}_1}{\partial y} \end{pmatrix} = \boldsymbol{J}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_1}{\partial r} \\ \frac{\partial \mathcal{N}_1}{\partial s} \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} c_y & -b_y \\ -c_x & b_x \end{pmatrix} \cdot \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} b_y - c_y \\ c_x - b_x \end{pmatrix}$$

$$\begin{pmatrix} \frac{\partial \mathcal{N}_2}{\partial x} \\ \frac{\partial \mathcal{N}_2}{\partial y} \end{pmatrix} = \boldsymbol{J}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_2}{\partial r} \\ \frac{\partial \mathcal{N}_2}{\partial s} \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} c_y & -b_y \\ -c_x & b_x \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} c_y \\ -c_x \end{pmatrix}$$

$$\begin{pmatrix} \frac{\partial \mathcal{N}_3}{\partial x} \\ \frac{\partial \mathcal{N}_3}{\partial y} \end{pmatrix} = \boldsymbol{J}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_3}{\partial r} \\ \frac{\partial \mathcal{N}_3}{\partial s} \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} c_y & -b_y \\ -c_x & b_x \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{2A} \begin{pmatrix} -b_y \\ b_x \end{pmatrix}$$

In the end this means that if a field ϕ is known at the vertices of the triangle (with values $\phi_{0,1,2}$) its value anywhere in the triangle is given by

$$\phi^h(x, y) = \sum_i \mathcal{N}_i(r(x, y), s(x, y))\phi_i$$

and its gradient by

$$\nabla^h \phi(x, y) = \sum_i \begin{pmatrix} \sum_i \frac{\partial \mathcal{N}_i}{\partial x} \phi_i \\ \sum_i \frac{\partial \mathcal{N}_i}{\partial y} \phi_i \end{pmatrix}$$

where we must note that the gradient is in fact constant over the whole triangle (it does not depend on x, y).

11.4   Bilinear mapping (Q_1) on a quadrilateral

The reference element is in the (r, s) space. It is a square of size 2×2 centered around the origin, i.e. $(r, s) \in [-1, 1] \times [-1, 1]$. We wish to map it to the quadrilateral in the (x, y) space (and vice versa):
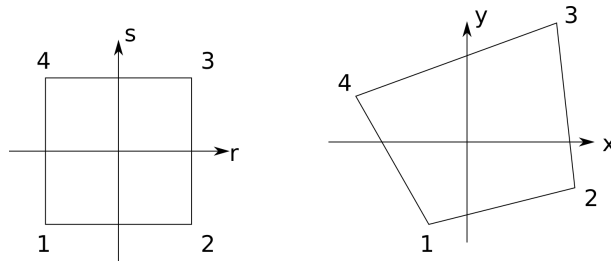
By convention we call *forward mapping* the mapping from (r, s) to (x, y).

The coordinates of the vertices are (x_1, y_1), (x_2, y_2), (x_3, y_3) and (x_4, y_4). We then simply have the following relationship, i.e. any point of the reference element can be mapped to the physical quadrilateral as follows:

$$\begin{aligned} x(r, s) &= \mathcal{N}_1(r, s)x_1 + \mathcal{N}_2(r, s)x_2 + \mathcal{N}_3(r, s)x_3 + \mathcal{N}_4(r, s)x_4 \\ y(r, s) &= \mathcal{N}_1(r, s)y_1 + \mathcal{N}_2(r, s)y_2 + \mathcal{N}_3(r, s)y_3 + \mathcal{N}_4(r, s)y_4 \end{aligned} \tag{11.12}$$

where the Q_1 basis functions $\mathcal{N}_i(r, s)$ are defined in Section 8.2.

In the following example the program randomly generates 10000 points inside the reference element and computes their mapping into the (x, y) space.

```
x1=-1  ;  y1=-2
x2=3   ;  y2=-1
x3=2   ;  y3=2
x4=-3  ;  y4=1

npts=10000
r=np.zeros(npts,dtype=np.float64)
s=np.zeros(npts,dtype=np.float64)
x=np.zeros(npts,dtype=np.float64)
y=np.zeros(npts,dtype=np.float64)

for i in range(0,npts):
    # compute random r,s coordinates
    r[i]=random.uniform(-1.,+1)
    s[i]=random.uniform(-1.,+1)
    # compute basis function values at r,s
    N1=0.25*(1-r[i])*(1-s[i])
    N2=0.25*(1+r[i])*(1-s[i])
    N3=0.25*(1+r[i])*(1+s[i])
    N4=0.25*(1-r[i])*(1+s[i])
    # compute x,y coordinates
    x[i]=N1*x1+N2*x2+N3*x3+N4*x4
    y[i]=N1*y1+N2*y2+N3*y3+N4*y4

np.savetxt('rs.ascii',np.array([r,s]).T)
np.savetxt('xy.ascii',np.array([x,y]).T)
```
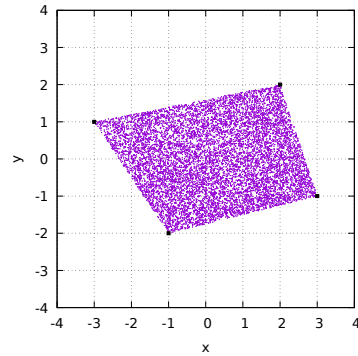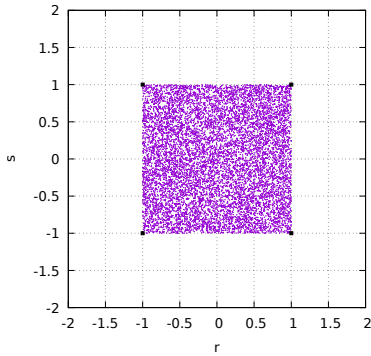

There is also an inverse map, which is not so easily computed (see Section 15.11). However, if the quadrilateral in the (x, y) space is a rectangle of size (h_x, h_y), the inverse mapping is trivial:

$$r \;=\; \frac{x - x_1}{x_2 - x_1} \tag{11.13}$$

$$s \;=\; \frac{y - y_1}{y_4 - y_1} \tag{11.14}$$

Also in the case of rectangular elements of size (h_x, h_y) the basis functions can directly be written as

533

functions of the (x, y) coordinates:

$$N_1(x, y) = \left(\frac{x_3 - x}{h_x}\right)\left(\frac{y_3 - y}{h_y}\right)$$

$$N_2(x, y) = \left(\frac{x - x_1}{h_x}\right)\left(\frac{y_3 - y}{h_y}\right)$$

$$N_3(x, y) = \left(\frac{x - x_1}{h_x}\right)\left(\frac{y - y_1}{h_y}\right)$$

$$N_4(x, y) = \left(\frac{x_3 - x}{h_x}\right)\left(\frac{y - y_1}{h_y}\right)$$

From Eq. (11.12) and using the expressions for the Q_1 basis functions, we can write

$$x = \frac{1}{4}(x_1 + x_2 + x_3 + x_4) + \frac{1}{4}(-x_1 + x_2 + x_3 - x_4)r + \frac{1}{4}(-x_1 - x_2 + x_3 + x_4)s + \frac{1}{4}(x_1 - x_2 + x_3 - x_4)rs$$

$$y = \frac{1}{4}(y_1 + y_2 + y_3 + y_4) + \frac{1}{4}(-y_1 + y_2 + y_3 - y_4)r + \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)s + \frac{1}{4}(y_1 - y_2 + y_3 - y_4)rs$$

Let us compute the space derivatives of these quantities:

$$\frac{\partial x}{\partial r} = \frac{1}{4}(-x_1 + x_2 + x_3 - x_4) + \frac{1}{4}(x_1 - x_2 + x_3 - x_4)s = A_1 + A_2 s$$

$$\frac{\partial x}{\partial s} = \frac{1}{4}(-x_1 - x_2 + x_3 + x_4) + \frac{1}{4}(x_1 - x_2 + x_3 - x_4)r = A_3 + A_4 r$$

$$\frac{\partial y}{\partial r} = \frac{1}{4}(-y_1 + y_2 + y_3 - y_4) + \frac{1}{4}(y_1 - y_2 + y_3 - y_4)s = B_1 + B_2 s$$

$$\frac{\partial y}{\partial s} = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) + \frac{1}{4}(y_1 - y_2 + y_3 - y_4)r = B_3 + B_4 r$$

The jacobian matrix is then given by

$$J = \left(\begin{array}{cc} A_1 + A_2 s & B_1 + B_2 s \\ A_3 + A_4 r & B_3 + B_4 r \end{array}\right)$$

and its inverse

$$J^{-1} = \frac{1}{C}\left(\begin{array}{cc} B_3 + B_4 r & -B_1 - B_2 s \\ -A_3 - A_4 r & A_1 + A_2 s \end{array}\right)$$

with C being the determinant given by

$$C = (A_1 + A_2 s)(B_3 + B_4 r) - (A_3 + A_4 r)(B_1 + B_2 s)$$

11.4.1 A concrete example

Let us look at this by means of a simple example and let us consider the following element:

534
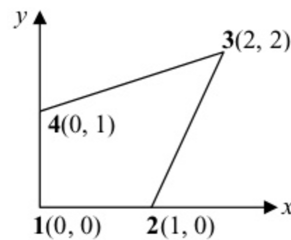
Then a Q_1 mapping yields:

$$
\begin{aligned}
x(r,s) &= \sum_i \mathcal{N}_i(r,s)x_i = \mathcal{N}_2 + 2\mathcal{N}_3 = \frac{1}{4}(3 + 3r + s + rt) \\
y(r,s) &= \sum_i \mathcal{N}_i(r,s)y_i = 2\mathcal{N}_3 + \mathcal{N}_4 = \frac{1}{4}(3 + r + 3s + rt)
\end{aligned}
\tag{11.15}
$$

The Jacobian matrix is then

$$
\boldsymbol{J} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\[2mm] \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 3+s & 1+s \\ 1+r & 3+r \end{pmatrix}
$$

and its determinant is

$$
|\boldsymbol{J}| = \frac{1}{4}[(3+s)(3+r) - (1+s)(1+r)] = \frac{1}{2} + \frac{1}{8}r + \frac{1}{8}s
\tag{11.16}
$$

It is clear that $|\boldsymbol{J}| > 0$ for $-1 \le r \le +1$ and $-1 \le s \le +1$ so that forward and inverse mappings exist and are unique.

Let us now consider another example, the following element:
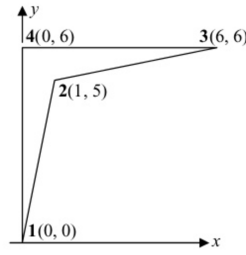
It follows that

$$
\begin{aligned}
x(r,s) &= \sum_i \mathcal{N}_i(r,s)x_i = \frac{1}{4}(1+r)(7+5s)
\end{aligned}
\tag{11.17}
$$

$$
\begin{aligned}
y(r,s) &= \sum_i \mathcal{N}_i(r,s)y_i = \frac{1}{4}(17 + 5r + 7s - 5rs)
\end{aligned}
\tag{11.18}
$$

and the determinant:

$$
|\boldsymbol{J}| = \frac{3}{2} - \frac{15r}{4} + \frac{15s}{4}
$$

is zero for $r - s = 2/5$. This mapping is invalid!

11.4.2 the special case of squares, quadrilaterals and parallelograms

In this case the parameters A_2, A_4, B_2, B_4 are all equal to zero. We can convince ourselves that it is indeed the case by drawing these quadrilaterals and computing these quantities.

535
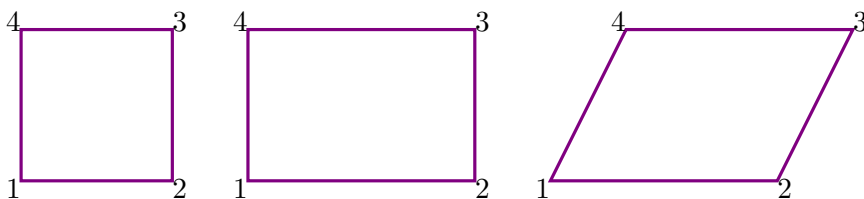
For the square of size h we have

$$
\begin{aligned}
4A_2 &= x_1 - x_2 + x_3 - x_4 = x_1 - (x_1 + h) + (x_4 + h) - x_4) = 0 \\
4A_4 &= x_1 - x_2 + x_3 - x_4 = x_1 - (x_1 + h) + (x_4 + h) - x_4 = 0 \\
4B_2 &= y_1 - y_2 + y_3 - y_4 = y_1 - y_2 + (y_2 + h_y) - (y_1 + h_y) = 0 \\
4B_4 &= y_1 - y_2 + y_3 - y_4 = y_1 - y_2 + (y_2 + h_y) - (y_1 + h_y) = 0
\end{aligned}
$$

For the rectangle of size h_x, h_y we have

$$
\begin{aligned}
4A_2 &= x_1 - x_2 + x_3 - x_4 = x_1 - (x_1 + h_x) + (x_4 + h_x) - x_4) = 0 \\
4A_4 &= x_1 - x_2 + x_3 - x_4 = x_1 - (x_1 + h_x) + (x_4 + h_x) - x_4 = 0 \\
4B_2 &= y_1 - y_2 + y_3 - y_4 = y_1 - y_2 + (y_2 + h_y) - (y_1 + h_y) = 0 \\
4B_4 &= y_1 - y_2 + y_3 - y_4 = y_1 - y_2 + (y_2 + h_y) - (y_1 + h_y) = 0
\end{aligned}
$$

and the same for the parallelogram.

In the case of a rectangle we also have

$$
\begin{aligned}
A_1 &= \frac{1}{4}(-x_1 + x_2 + x_3 - x_4) = \frac{1}{4}(-x_1 + (x_1 + h_x) + (x_4 + h_x) - x_4) = \frac{h_x}{2} \\
A_3 &= \frac{1}{4}(-x_1 - x_2 + x_3 + x_4) = \frac{1}{4}(-x_1 - (x_1 + h_x) + (x_4 + h_x) + x_4) = 0 \\
B_1 &= \frac{1}{4}(-y_1 + y_2 + y_3 - y_4) = \frac{1}{4}(-y_1 + y_2 + (y_2 + h_y) - (y_1 + h_y)) = 0 \\
B_3 &= \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) = \frac{1}{4}(-y_1 - y_2 + (y_2 + h_y) + (y_1 + h_y)) = \frac{h_y}{2}
\end{aligned}
$$

and something very similar for the square with $h = h_x = h_y$.

For these quadrilaterals we then have

$$
\boldsymbol{J}^{-1} = \frac{1}{C}\begin{pmatrix} B_3 & -B_1 \\ -A_3 & A_1 \end{pmatrix} \qquad \text{with} \qquad C = A_1 B_3 - A_3 B_1
$$

and in the case of the rectangle:

$$
\boldsymbol{J} = \begin{pmatrix} A_1 & B_1 \\ A_3 & B_3 \end{pmatrix} = \begin{pmatrix} \frac{h_x}{2} & 0 \\ 0 & \frac{h_y}{2} \end{pmatrix}
$$

The determinant is then $C = \frac{h_x h_y}{4}$ and the inverse:

$$
\boldsymbol{J}^{-1} = \frac{1}{C}\begin{pmatrix} B_3 & -B_1 \\ -A_3 & A_1 \end{pmatrix} = \frac{4}{h_x h_y}\begin{pmatrix} \frac{h_y}{2} & 0 \\ 0 & \frac{h_x}{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{h_x} & 0 \\ 0 & \frac{2}{h_y} \end{pmatrix}
$$

Remark. *Hua [632] (1990) has published analytical inverse transformation for quadrilateral isoparametric elements, i.e. how to compute $\boldsymbol{J}^{-1}$ as a function of space coordinates and not just at the quadrature points.*

11.5 Biquadratic mapping of a straight-edge face Q_2 element
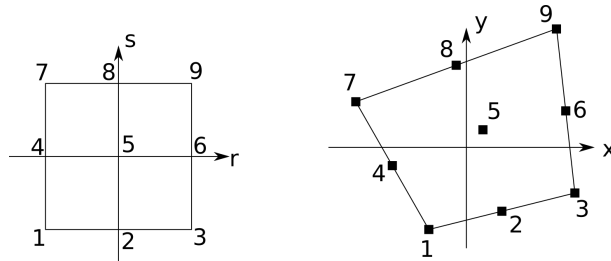
The reference element now contains 9 nodes: 1,3,7,9 are the corners, nodes 2,4,6,8 are the mid-face points and node 5 is in the middle[3]. The mapping from the (r, s) space to the (x, y) space is then as follows:

$$\begin{pmatrix} x(r,s) \\ y(r,s) \end{pmatrix} = \sum_{i=1}^{9} \mathcal{N}_i(r,s) \begin{pmatrix} x_i \\ y_i \end{pmatrix} = \mathcal{N}_1(r,s) \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathcal{N}_2(r,s) \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} + \mathcal{N}_3(r,s) \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}$$
$$+ \mathcal{N}_4(r,s) \begin{pmatrix} x_4 \\ y_4 \end{pmatrix} + \mathcal{N}_5(r,s) \begin{pmatrix} x_5 \\ y_5 \end{pmatrix} + \mathcal{N}_6(r,s) \begin{pmatrix} x_6 \\ y_6 \end{pmatrix}$$
$$+ \mathcal{N}_7(r,s) \begin{pmatrix} x_7 \\ y_7 \end{pmatrix} + \mathcal{N}_8(r,s) \begin{pmatrix} x_8 \\ y_8 \end{pmatrix} + \mathcal{N}_9(r,s) \begin{pmatrix} x_9 \\ y_9 \end{pmatrix}$$

where the Q_2 basis functions have been obtained in Section 8.3.2:

$$\begin{aligned}
\mathcal{N}_1(r,t) &= 0.5r(r-1)0.5t(t-1) \\
\mathcal{N}_2(r,t) &= (1-r^2)0.5t(t-1) \\
\mathcal{N}_3(r,t) &= 0.5r(r+1)0.5t(t-1) \\
\mathcal{N}_4(r,t) &= 0.5r(r-1)(1-t^2) \\
\mathcal{N}_5(r,t) &= (1-r^2)(1-t^2) \\
\mathcal{N}_6(r,t) &= 0.5r(r+1)(1-t^2) \\
\mathcal{N}_7(r,t) &= 0.5r(r-1)0.5t(t+1) \\
\mathcal{N}_8(r,t) &= (1-r^2)0.5t(t+1) \\
\mathcal{N}_9(r,t) &= 0.5r(r+1)0.5t(t+1)
\end{aligned}$$

As in the previous section I have written a simple Python script that computes the image of 10,000 points generated in the reference element using the forward mapping based on the Q_2 basis functions:

```
x1=−1                    ;  y1=−2
x3=3                     ;  y3=−1
x9=2                     ;  y9=2
x7=−3                    ;  y7=1
x2=0.5*(x1+x3)           ;  y2=0.5*(y1+y3)
x4=0.5*(x1+x7)           ;  y4=0.5*(y1+y7)
x6=0.5*(x3+x9)           ;  y6=0.5*(y3+y9)
x8=0.5*(x7+x9)           ;  y8=0.5*(y7+y9)
x5=0.25*(x1+x3+x7+x9)   ;  y5=0.25*(y1+y3+y7+y9)


npts=10000
r=np.zeros(npts,dtype=np.float64)
s=np.zeros(npts,dtype=np.float64)
xQ1=np.zeros(npts,dtype=np.float64)
yQ1=np.zeros(npts,dtype=np.float64)
xQ2=np.zeros(npts,dtype=np.float64)
yQ2=np.zeros(npts,dtype=np.float64)


for i in range(0,npts):
    # compute random r,s coordinates
    r[i]=random.uniform(−1.,+1)
    s[i]=random.uniform(−1.,+1)
    # compute Q2 basis function values at r,s
    N1= 0.5*r[i]*(r[i]−1.) * 0.5*s[i]*(s[i]−1.)
```

[3] Note that this numbering is quite arbitrary

```
    N2=            (1.−r[i]**2)  *  0.5*s[i]*(s[i]−1.)
    N3=  0.5*r[i]*(r[i]+1.)  *  0.5*s[i]*(s[i]−1.)
    N4=  0.5*r[i]*(r[i]−1.)  *          (1.−s[i]**2)
    N5=            (1.−r[i]**2)  *          (1.−s[i]**2)
    N6=  0.5*r[i]*(r[i]+1.)  *          (1.−s[i]**2)
    N7=  0.5*r[i]*(r[i]−1.)  *  0.5*s[i]*(s[i]+1.)
    N8=            (1.−r[i]**2)  *  0.5*s[i]*(s[i]+1.)
    N9=  0.5*r[i]*(r[i]+1.)  *  0.5*s[i]*(s[i]+1.)
    # compute x,y coordinates
    xQ2[i]=N1*x1+N2*x2+N3*x3+N4*x4+N5*x5+N6*x6+N7*x7+N8*x8+N9*x9
    yQ2[i]=N1*y1+N2*y2+N3*y3+N4*y4+N5*y5+N6*y6+N7*y7+N8*y8+N9*y9
    # compute Q1 basis function values at r,s
    N1=0.25*(1−r[i])*(1−s[i])
    N2=0.25*(1+r[i])*(1−s[i])
    N3=0.25*(1+r[i])*(1+s[i])
    N4=0.25*(1−r[i])*(1+s[i])
    # compute x,y coordinates
    xQ1[i]=N1*x1+N2*x3+N3*x9+N4*x7
    yQ1[i]=N1*y1+N2*y3+N3*y9+N4*y7

np.savetxt('rs.ascii',np.array([r,s]).T)
np.savetxt('xyQ1.ascii',np.array([xQ1,yQ1]).T)
np.savetxt('xyQ2.ascii',np.array([xQ2,yQ2]).T)
```

The code is available in `/images/mappings/biquadratic` Note that the coordinates of point 5 are defined being those of the average of the corner coordinates of the quadrilateral. More on this choice later.
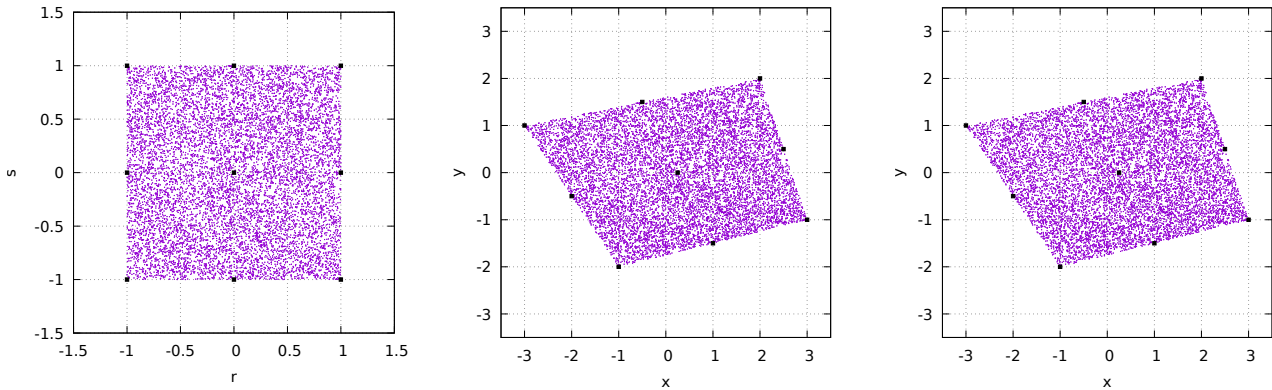
Left: 10,000 random points in the reference element; image of these points by means of a bilinear (middle) and biquadratic (right) mapping respectively.

When the sides of the element are straight we see that a Q_1 mapping is sufficient.

11.6 Biquadratic mapping of a not-so straight-line face Q_2 element

We now carry out the same exercise as before but nodes 2 and 8 are no more in the middle of nodes 1-3 and 7-9 respectively. The code is available in `/images/mappings/biquadratic2`.
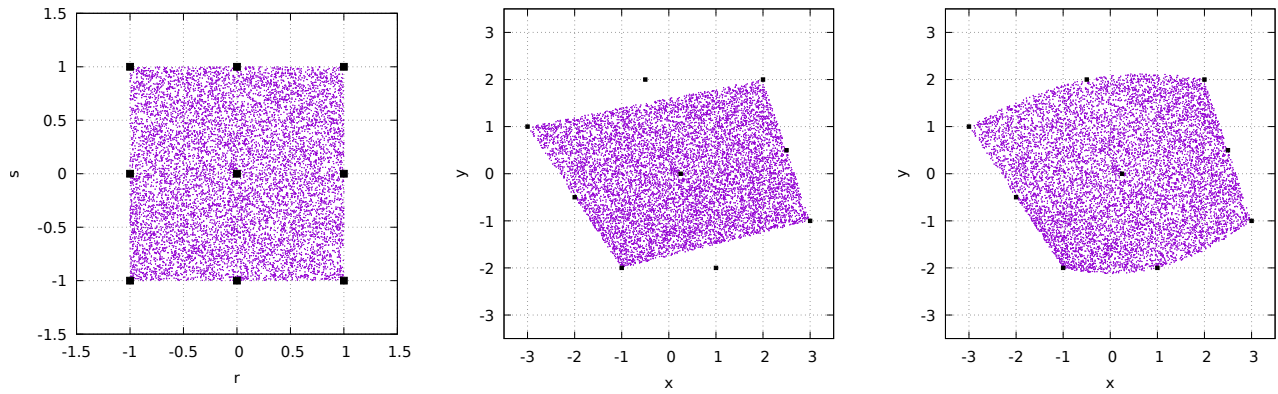
Left: 10,000 random points in the reference element; image of these points by means of a bilinear (middle) and biquadratic mapping respectively (right).

In this case we see that the Q_2 mapping manages to better capture the 'real' shape of the element. Since nodes 2 and 8 have moved, we could now ask ourselves where we should place node 5? In this example we set it as follows but it is somewhat arbitrary.

```
x5=(x1+x2+x3+x4+x6+x7+x8+x9)/8.
y5=(y1+y2+y3+y4+y6+y7+y8+y9)/8.
```

We will come back to this later.

11.7 Q_1, Q_2, Q_3 and Q_4 mappings in an annulus

In the light of what precedes, we can now ask ourselves how this translates to a real geodynamic case. Let us then consider the case of an annular domain, a cross section of a hollow sphere. When using quadrilateral elements, the mesh will look similar to this:
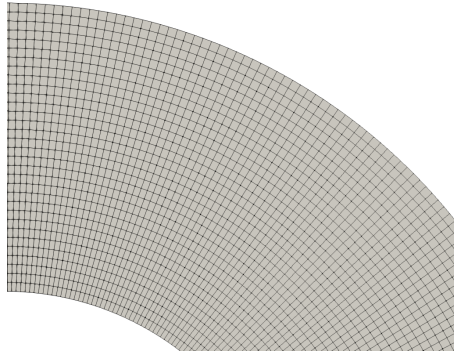
We here focus on Q_1, Q_2, Q_3 and Q_4 mappings. We single out an element, and arbitrarily define it as follows in polar coordinates:

```
theta1=23./180.*np.pi
theta2=52./180.*np.pi
R1=1.
R2=1.5
```

The Q_1 mapping requires four points, the Q_2 nine points, the Q_3 sixteen nodes and the Q_4 twenty-five points. The code used in the following is available at ./images/mappings/curved/. These are placed equidistantly in the r, θ coordinate system, as shown hereunder:

Position of the nodes for the Q_1, Q_2, Q_3 and Q_4 mappings.
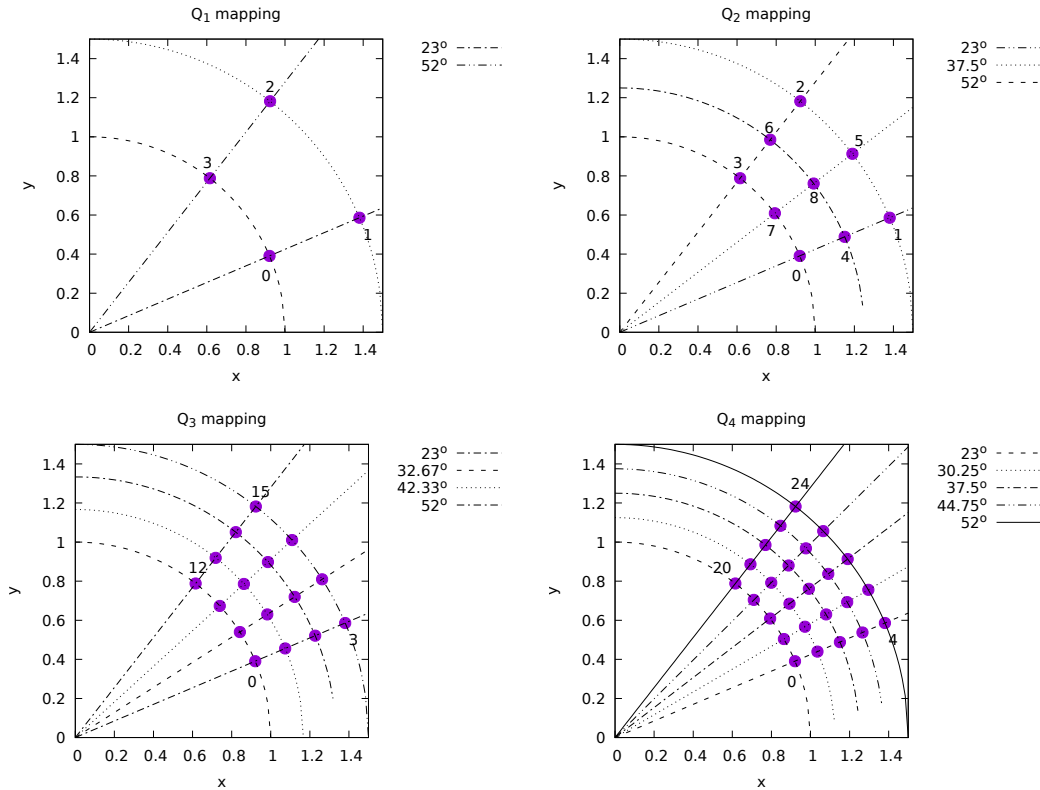
As before, we randomly shoot 10,000 points inside the reference element and map these out in the x, y space. Resulting swarms of points are shown in the following figures:

540
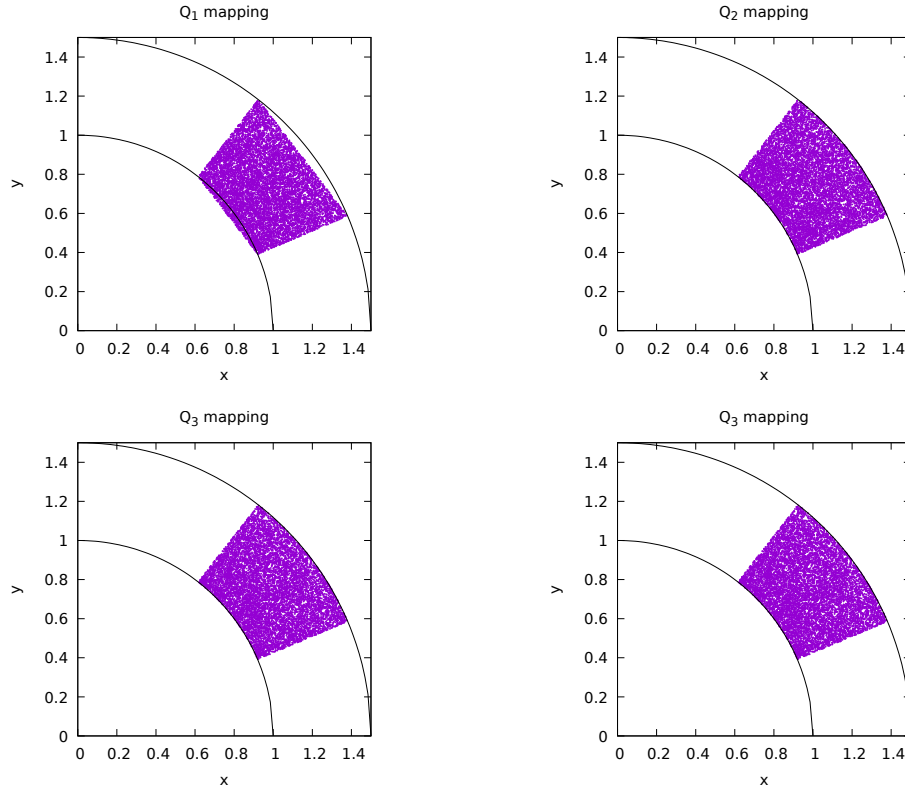
Left to right: position of the mapped points for the Q_1, Q_2, Q_3 and Q_4 mappings.

The image of a square with a Q_1 mapping is obviously a quadrilateral with straight edges so that it looks like quite a few points land outside of the domain $R_1 \leq r \leq R_2$. Note that points are well within $23° \leq \theta \leq 52°$, which can simply be explained by the fact that the faces of the element joining R_1 to R_2 are straight lines.

However, it looks like the higher order mappings are doing a much better job at mapping the region of space $R_1 \leq r \leq R_2$. In order to characterise this better, we now place 10,000 points on the bottom face of the reference element (i.e. $s = -1$) and once again compute their coordinates in the the x, y space:
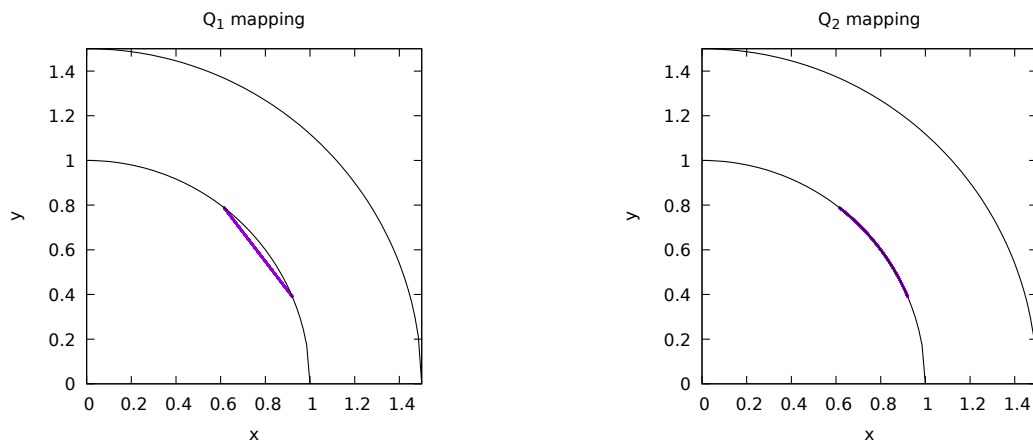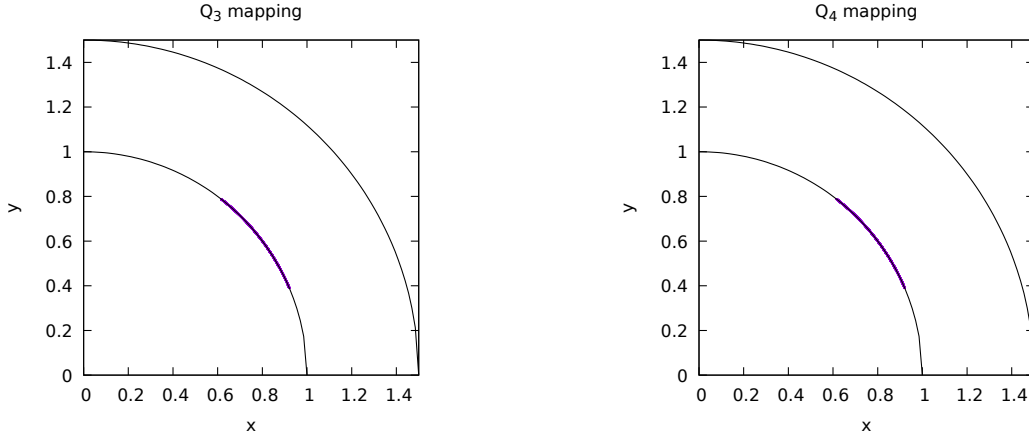
Position of the mapped points for the Q_1, Q_2, Q_3 and Q_4 mappings.

For each point i we now compute the distance r_i to the origin, which, if the mapping was perfect, would be exactly equal to $R_1 = 1$. On the following plots are shown the error $r_i - 1$ for all points, from $r = -1$ to $r = +1$.
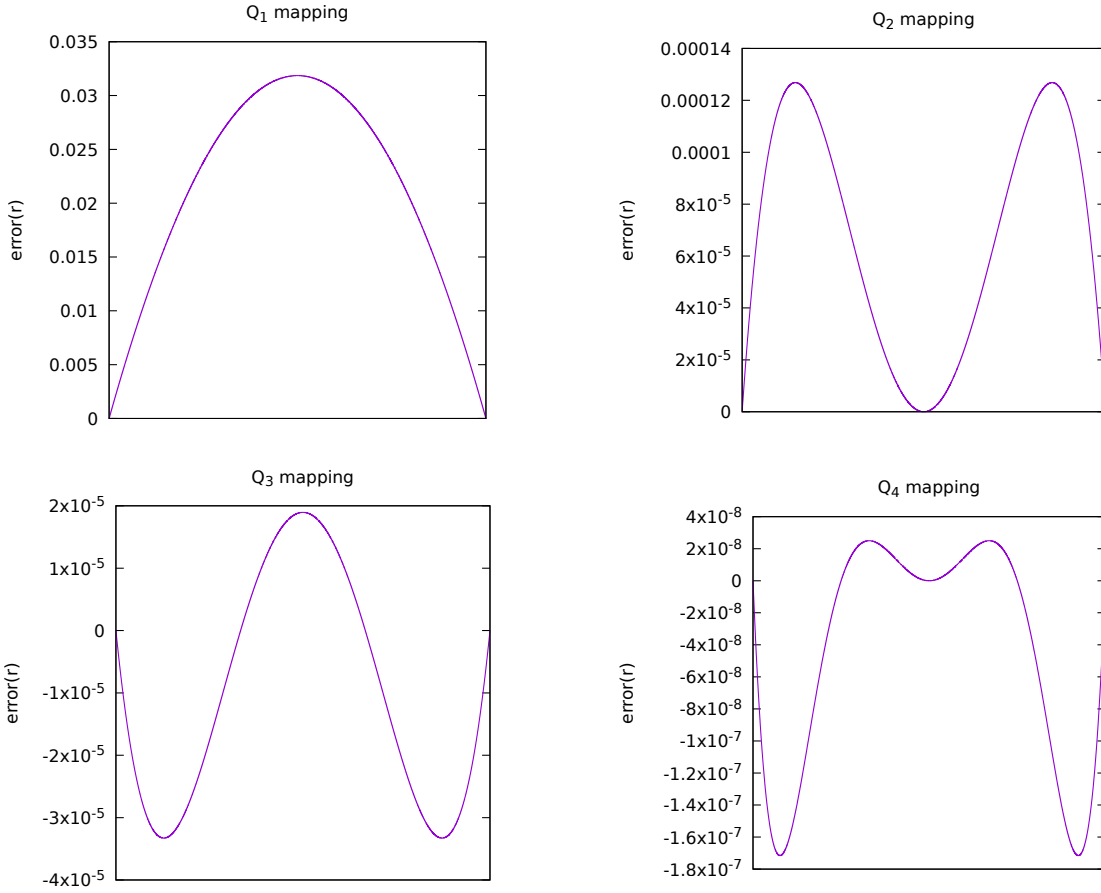
Radius error of the mapped points for the Q_1, Q_2, Q_3 and Q_4 mappings.

We see that the amplitude of the error decreases with the order of the mapping used, which is why for instance ASPECT uses a Q_4 mapping by default[4]. Actually, in this particular case, the equation which describes the circle is not a polynomial so that no high-order mapping will ever be able to *exactly* represent the curved boundary of the element!

---

[4]I find it also quite striking that the Q_4 mapping outperforms the Q_3 one by two orders of magnitude...

Another interesting point to keep in mind is that the location of the quadrature points in the x, y space is also determined by the mapping used, which can have consequences on the accuracy of the integration and it will be reflected (for instance) on the error convergence rate.

As already mentioned previously, the coordinates of the nodes of the element in the x, y are uniquely determined when they are on the convex hull of the element (for instance nodes 0-7 for Q_2) but we need to choose the position of the last nodes which are inside the element. Unfortunately, this choice is not neutral.

Finally, we can explore the importance of the mapping in combination with numerical quadrature. For each mapping we compute the area of the element by means of a 3x3, 4x4 or 5x5 quadrature.

```
**********Q1*********
nqperdim= 3 0.3030060126539606 rel. error -0.04215361698430029
nqperdim= 4 0.3030060126539606 rel. error -0.04215361698430012 ~ 4%
nqperdim= 5 0.3030060126539606 rel. error -0.04215361698430012
**********Q2*********
nqperdim= 3 0.3162980025394154 rel. error -0.00013569026611326453
nqperdim= 4 0.3162980025394155 rel. error -0.00013569026611308905 ~ 0.01%
nqperdim= 5 0.3162980025394154 rel. error -0.00013569026611326453
**********Q3*********
nqperdim= 3 0.3163472223929359 rel. error 1.9900899402587318e-05
nqperdim= 4 0.316347222392936  rel. error 1.9900899402938278e-05 ~ 0.002%
nqperdim= 5 0.316347222392936  rel. error 1.9900899402938278e-05
**********Q4*********
nqperdim= 3 0.3163409410866220 rel. error 4.477021014282521e-08
nqperdim= 4 0.3163409541901677 rel. error 8.619243716974044e-08 ~ 0.000008%
nqperdim= 5 0.316340954190168  rel. error 8.619243804713484e-08
```

Here again the Q_4 mapping makes quite the difference.

11.8 P_1, P_2, P_3 and P_4 mappings in an annulus

The same exercise as in the previous section can be carried with triangular elements. Please read the previous section to understand the following plots. The code used in the following is available at ./images/mappings/curved_triangles/.

The major difference with respect to the quadrilateral element of the previous section is the location of the nodes for $P_{3,4}$. Indeed for quadrilaterals it made sense to place the nodes on rows at a constant radius and columns at a constant angle. Looking at P_2 the situation is simple: node 3 is the middle of the edge 0-1, node 5 is the middle of edge 0-2, and it seems reasonable to place node 4 at a position characterized by $r_4 = R_2 = 1.5$ and $\theta_4 = \frac{1}{2}(\theta_1 + \theta_2)$. Concerning P_3 and P_4 I simply use a P_1 forward mapping for all 10 and 15 nodes respectively. I then proceed to move nodes 6,8 and 8,11,13 respectively so that they are placed on the circle at equi-angular distance for one another[5].

Position of the nodes for the P_1, P_2, P_3 and P_4 mappings.
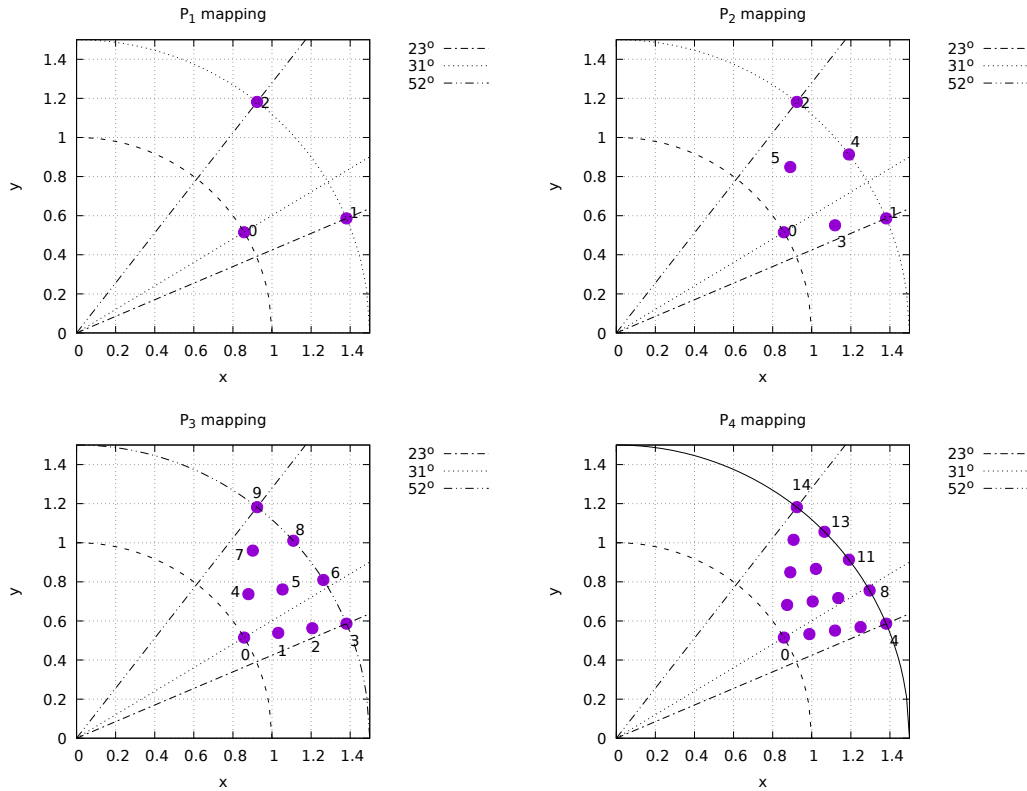
Once the nodes have been layed out we can then use the $P_{1,2,3,4}$ to map out 10,000 points:

---
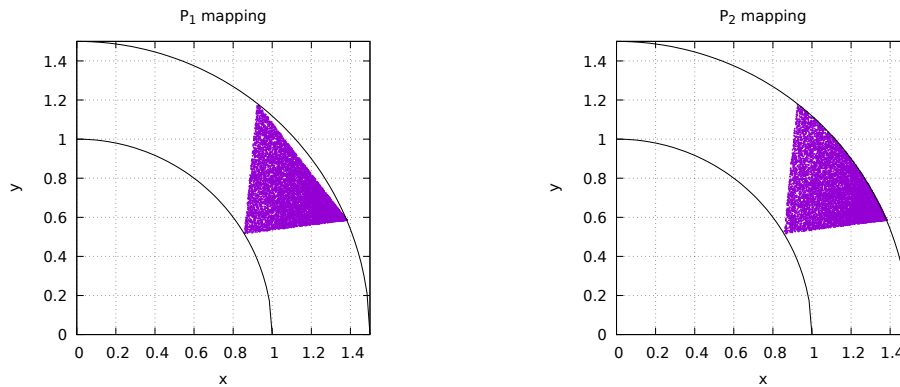
[5]This is obviously very arbitrary and could/should be improved but how?
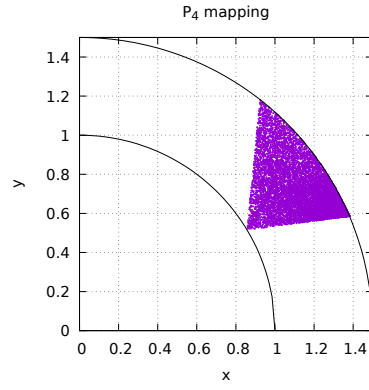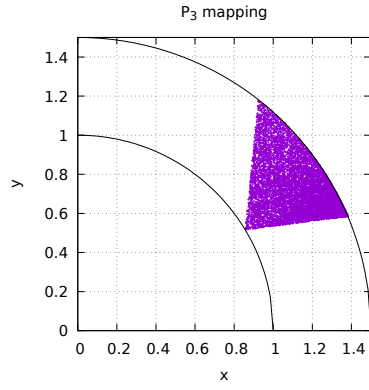
Left to right: position of the mapped points for the P_1, P_2, P_3 and P_4 mappings.

We now place 10,000 points on the face of the reference element that is mapped onto the circle and once again compute their coordinates in the the x, y space:
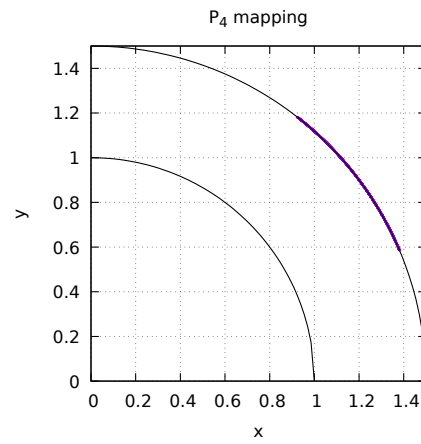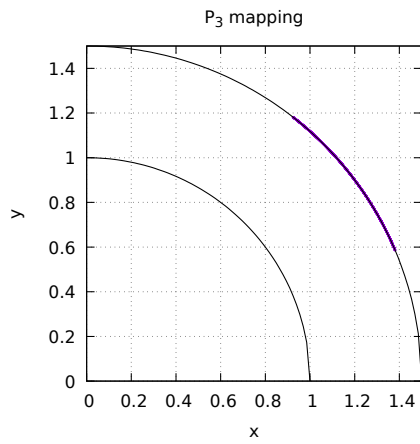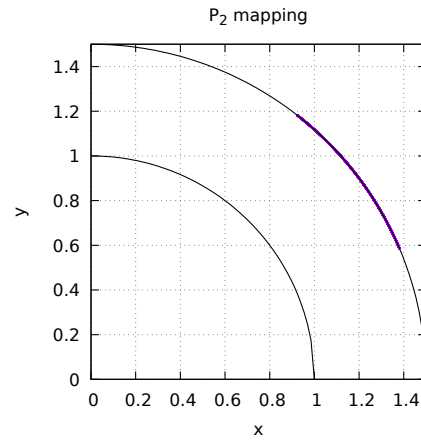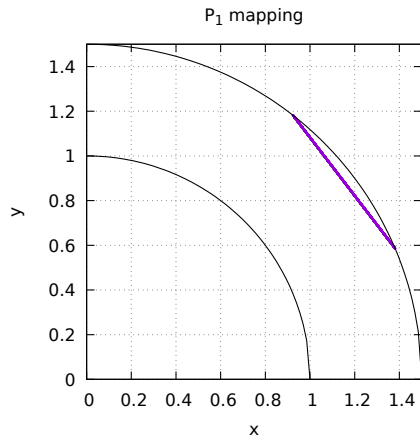
Position of the mapped points for the P_1, P_2, P_3 and P_4 mappings.

For each point i we now compute the distance r_i to the origin, which, if the mapping was perfect, would be exactly equal to $R_2 = 1.5$. On the following plots are shown the error $r_i - R_2$ for all points:
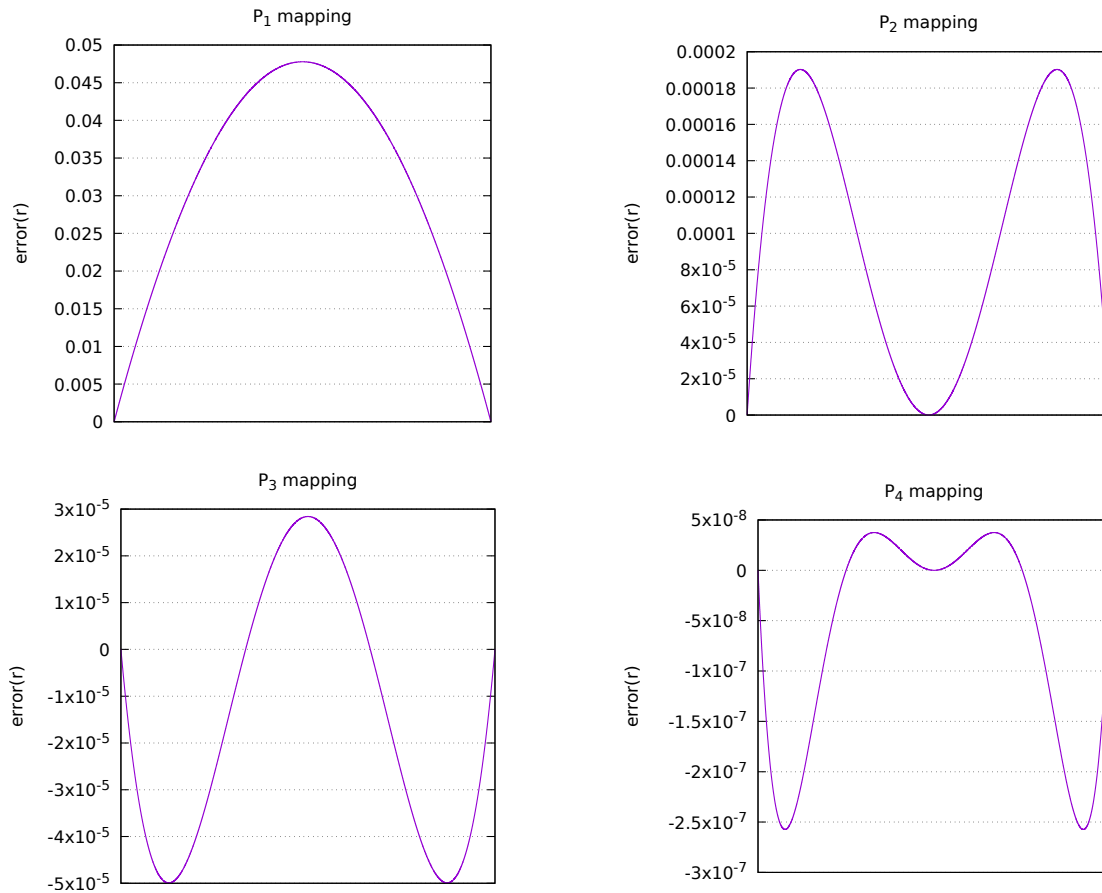
Radius error of the mapped points for the P_1, P_2, P_3 and P_4 mappings.

We recover the same kind of plots as with the quadrilaterals.

When it comes to integration the problem we are facing here is the fact that the analytical value of the area covered by the shape formed by the 2 straight edges and the curved side is very difficult to compute.

```
**********P1*********
nqel= 3 area= 0.17225503489810473
nqel= 4 area= 0.1722550348981049
nqel= 6 area= 0.17225503489810473
nqel= 7 area= 0.172255034898036
**********P2*********
nqel= 3 area= 0.19618061669192358
nqel= 4 area= 0.19618061669192355
nqel= 6 area= 0.19618061669192338
nqel= 7 area= 0.19618061669184547
**********P3*********
nqel= 3 area= 0.196202089783442
nqel= 4 area= 0.19620751672068115
nqel= 6 area= 0.19626921242826045
nqel= 7 area= 0.1962692124281841
**********P4*********
nqel= 3 area= 0.19339109577274016
nqel= 4 area= 0.1961927497201587
nqel= 6 area= 0.1962431894093658
nqel= 7 area= 0.19626032062958276
```

11.9 Biquadratic mapping - the middle node conundrum

Python code at `images/mappings/biquadratic3`.

As mentioned before, unless the element is a straight-edge quadrilateral, determining the (best) position of the middle node is not trivial. Or is it?

Let us recall the internal numbering of nodes:

```
4--7--3
|     |
8  9  6    (reference element)
|     |
1--5--2
```

We can think of multiple ways to come up with the 'center' of the element, i.e. the location of node #9. In the code these different approaches are parameterized by means of the center variable:

- center=0: The middle node coordinates are simply the average of the corner coordinates

$$x_9 = (x_1 + x_2 + x_3 + x_4)/4$$
$$y_9 = (y_1 + y_2 + y_3 + y_4)/4$$

- center=1: The middle coordinates are the average all the 4 corners and 4 mid-edge nodes coordinates:

$$x_9 = (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8)/8$$
$$y_9 = (y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8)/8$$

- center=2: The middle coordinates are obtained via a weighed sum of the corner coordinates and the mid-edge nodes coordinates[6]:

$$x_9 = (x_1 + x_2 + x_3 + x_4 + 3x_5 + 3x_6 + 3x_7 + 3x_8)/16.$$
$$y_9 = (y_1 + y_2 + y_3 + y_4 + 3y_5 + 3y_6 + 3y_7 + 3y_8)/16.$$

- center=3: This only applies to element #4 that is defined on an annulus:

$$x_9 = \frac{1}{2}(R_1 + R_2)\cos(3\pi/8)$$
$$y_9 = \frac{1}{2}(R_1 + R_2)\sin(3\pi/8)$$

- center=4: This only applies to element #4 that is defined by $R_1 < r < R_2$ and $\theta_1 < \theta < \theta_2$. Node #9 is the center of mass.

> other elements

---

[6]I have no idea where this comes from.

We need to compute[7]

$$
\begin{aligned}
\vec{R} &= \frac{1}{M}\int \vec{r}\rho(\vec{r})dV \\
&= \frac{1}{M}\rho_0 \int \vec{r}dV \\
&= \frac{1}{M}\frac{M}{V}\int \vec{r}dV \\
&= \frac{1}{V}\int \vec{r}dV \\
&= \frac{1}{V}\int \begin{pmatrix} x \\ y \end{pmatrix} dV \\
&= \frac{1}{V}\int \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix} dV \\
&= \frac{1}{V}\int_{R_1}^{R_2}\int_{\theta_1}^{\theta_2} \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix} r\,dr\,d\theta \\
&= \frac{1}{\frac{1}{2}(R_2^2 - R_1^2)(\theta_2 - \theta_1)}\frac{1}{3}(R_2^3 - R_1^3)\begin{pmatrix} \sin\theta_2 - \sin\theta_1 \\ -\cos\theta_2 + \cos\theta_1 \end{pmatrix} \\
&\simeq \begin{pmatrix} 0.5801028000103104 \\ 1.4004920473554983 \end{pmatrix}
\end{aligned}
\tag{11.19}
$$

which corresponds to $r = 1.5158816686291174$ and $\theta = 67.5^o = 3\pi/8$.

- center=5: variable position

- center=6: This approach is a bit more complex than the previous ones. We start by acknowledging that we know the coordinates of nodes 1-8 in the reference space (r,s) and in the real space (x,y) so one could then use a $Q_2^{(8)}$ mapping (i.e. 'serendipity' Q_2) to place node #9.

  In this case the basis functions for this mapping are defined in Section 8.3.5:

$$
\begin{aligned}
\mathcal{N}_0(r,s) &= \frac{1}{4}(1-r)(1-s)(-r-s-1) \\
\mathcal{N}_1(r,s) &= \frac{1}{4}(1+r)(1-s)(r-s-1) \\
\mathcal{N}_2(r,s) &= \frac{1}{4}(1+r)(1+s)(r+s-1) \\
\mathcal{N}_3(r,s) &= \frac{1}{4}(1-r)(1+s)(-r+s-1) \\
\mathcal{N}_4(r,s) &= \frac{1}{2}(1-r^2)(1-s) \\
\mathcal{N}_5(r,s) &= \frac{1}{2}(1+r)(1-s^2) \\
\mathcal{N}_6(r,s) &= \frac{1}{2}(1-r^2)(1+s) \\
\mathcal{N}_7(r,s) &= \frac{1}{2}(1-r)(1-s^2)
\end{aligned}
$$

---

[7] https://en.wikipedia.org/wiki/Center_of_mass

We would then compute the location of node 9 as follows (its r, s coordinates are $0, 0$):

$$
\begin{aligned}
x_9 &= \sum_{i=1}^{8} \mathcal{N}_i(r_9, s_9) x_i \\
&= \sum_{i=1}^{8} \mathcal{N}_i(0, 0) x_i \\
&= -\frac{1}{4}(x_1 + x_2 + x_3 + x_4) + \frac{1}{2}(x_5 + x_6 + x_7 + x_8) \\
y_9 &= \sum_{i=1}^{8} \mathcal{N}_i(r_9, s_9) y_i \\
&= \sum_{i=1}^{8} \mathcal{N}_i(0, 0) y_i \\
&= -\frac{1}{4}(y_1 + y_2 + y_3 + y_4) + \frac{1}{2}(y_5 + y_6 + y_7 + y_8)
\end{aligned}
$$

About the code

In all what follows we use a Q_2 isoparametric mapping.

For simplicity I use in what follows the function $f(x, y) = x^3 + y^3$. At each point $(r, s) \in [-1, 1]^2$ with real coordinates $(x(r, s), y(r, s))$ we compute the error $e_f = \sum_i \mathcal{N}_i(r, s) f_i - f(xi(r, s), y(r, s))$ where f_i are the values of the function evaluated at the 9 nodes. We also compute the error on the first order derivatives, i.e. $e_{df/dx} = \sum_i \mathcal{N}_i(r, s)(df/dx)_i - df/dx(x(r, s), y(r, s))$ and $e_{df/dy} = \sum_i \mathcal{N}_i(r, s)(df/dy)_i - df/dy(x(r, s), y(r, s))$.

Unless indicated otherwise 3×3 quadrature is used to a) compute the area of the element; b) compute the errors above in the L_2 norm, e.g. $E_f = \left(\int_\Omega (e_f)^2 dV \right)^{1/2}$ (I purposefully do not divide by the area) and $E_{\nabla f} = \left(\int_\Omega (e_{df/dx}^2 + e_{df/dy}^2) dV \right)^{1/2}$.

We find that for all element shapes the measured area seems to be independent of the position of middle node. Why?

We note that the position of edges obtained from the forward mapping by setting $r = \pm 1$ and/or $s = \pm 1$ are in fact independent of the position of the middle node since shape functions are zero there. For example, setting $s = -1$ means that for terms containing this summation $\sum_i \mathcal{N}_i(r, s)...$ only contributions by nodes 1,5,2 count (basis functions $\mathcal{N}_{3,4,6,7,8,9}$ are zero for $s = -1$). When computing $y = \sum_i \mathcal{N}_i(r, s = -1) y_i$ then the result is independent of y_9. Same for all sides.

The Python code is built as follows:

```
for center in centers:
    assign middle point coordinates
    generate npts points at random in reference cell
    compute error on f,df/dx,df/dy and jacobian at each point
    export points and their fields to vtu
    compute area and integral of errors over element
```

Warning: Note that all the conclusions drawn hereafter are provisory: only *one* function was used which is arbitrarily chosen to be a 3rd order polynomial, and the coordinates/shapes of the elements 1,2,3 are also arbitrary and results may differ with different configurations!

Element 0
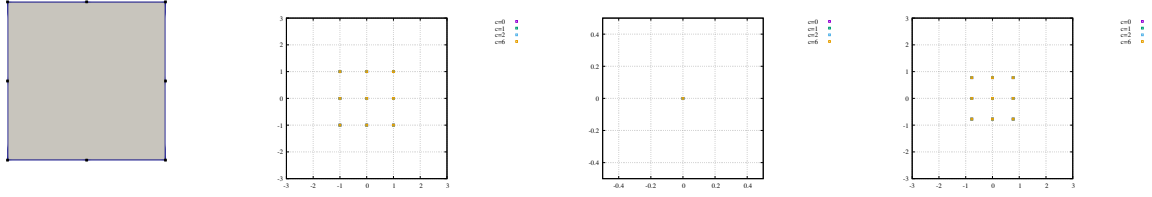
From left to right: element and its 8 boundary nodes. Location of the 9 nodes for different values of the center parameter. Zoom on the middle node. Position
of the quadrature points for different values of the center parameter.

x1=−1	;	y1=−1
x2=1	;	y2=−1
x3=1	;	y3=1
x4=−1	;	y4=1
x5=0.5*(x1+x2)	;	y5=0.5*(y1+y2)
x6=0.5*(x2+x3)	;	y6=0.5*(y2+y3)
x7=0.5*(x3+x4)	;	y7=0.5*(y3+y4)
x8=0.5*(x1+x4)	;	y8=0.5*(y1+y4)

In this case all only center=0,1,2,4,6 are applicable but they all lead to the same middle point
with $x_9 = 0, y_9 = 0$. This means that the position of quadrature points is also independent of the
center parameter.

```
element= 0, npts= 1000000, nqperdim= 3
============================================================
center= 0 x9,y9= 0.000000e+00 0.000000e+00
center= 0 min/max/avrg jcob on pts    = 1.000000e+00 1.000000e+00 1.000000e+00
center= 0 min/max/avrg err_f on pts   = -7.697997e-01 7.697981e-01 1.032961e-04
center= 0 min/max/avrg err_dfdx on pts=-1.999994e+00 1.000000e+00 9.897925e-04
center= 0 min/max/avrg err_dfdy on pts=-1.999998e+00 1.000000e+00 -4.337008e-04
center= 0 int err_f= 6.531973e-01
center= 0 int err_gradf= 2.529822e+00
center= 0 area= 4.000000e+00
============================================================
center= 1 x9,y9= 0.000000e+00 0.000000e+00
center= 1 min/max/avrg jcob on pts    = 1.000000e+00 1.000000e+00 1.000000e+00
center= 1 min/max/avrg err_f on pts   = -7.697981e-01 7.697978e-01 -1.037079e-04
center= 1 min/max/avrg err_dfdx on pts=-1.999994e+00 1.000000e+00 5.562081e-04
center= 1 min/max/avrg err_dfdy on pts=-1.999992e+00 1.000000e+00 -6.057413e-04
center= 1 int err_f= 6.531973e-01
center= 1 int err_gradf= 2.529822e+00
center= 1 area= 4.000000e+00
============================================================
center= 2 x9,y9= 0.000000e+00 0.000000e+00
center= 2 min/max/avrg jcob on pts    = 1.000000e+00 1.000000e+00 1.000000e+00
center= 2 min/max/avrg err_f on pts   = -7.698001e-01 7.697999e-01 -4.504134e-04
center= 2 min/max/avrg err_dfdx on pts=-1.999985e+00 1.000000e+00 4.678516e-04
center= 2 min/max/avrg err_dfdy on pts=-1.999987e+00 1.000000e+00 1.466204e-03
center= 2 int err_f= 6.531973e-01
center= 2 int err_gradf= 2.529822e+00
center= 2 area= 4.000000e+00
============================================================
center= 6 x9,y9= 0.000000e+00 0.000000e+00
center= 6 min/max/avrg jcob on pts    = 1.000000e+00 1.000000e+00 1.000000e+00
center= 6 min/max/avrg err_f on pts   = -7.697964e-01 7.698001e-01 -3.862548e-04
center= 6 min/max/avrg err_dfdx on pts=-1.999996e+00 1.000000e+00 -3.443542e-04
center= 6 min/max/avrg err_dfdy on pts=-1.999992e+00 1.000000e+00 -9.219252e-04
center= 6 int err_f= 6.531973e-01
center= 6 int err_gradf= 2.529822e+00
center= 6 area= 4.000000e+00
============================================================
```

We find that the jacobian is exactly 1 everywhere in the element which is what we expected since
$|J| = 4/h_x h_y = 4/2 \cdot 2 = 1$. Likewise area measurements are all identical and equal to $h_x h_y$.
Conclusions:

- In the case of square/rectangular elements the middle node location is indeed 'in the middle'.

- we obviously recover the same values for e_f and $E_{\nabla f}$

- Even using 1,000,000 points we find that the min/max/avrg values of pointwise errors show
  significant differences (esp. the average). In the future we'll then only consider E errors.

Element 1
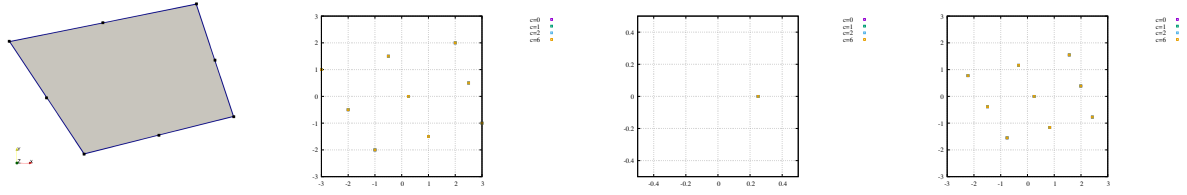
x1=−1	;	y1=−2
x2=3	;	y2=−1
x3=2	;	y3=2
x4=−3	;	y4=1
x5=0.5∗(x1+x2)	;	y5=0.5∗(y1+y2)
x6=0.5∗(x2+x3)	;	y6=0.5∗(y2+y3)
x7=0.5∗(x3+x4)	;	y7=0.5∗(y3+y4)
x8=0.5∗(x1+x4)	;	y8=0.5∗(y1+y4)

In this case only center=0,1,2,4,6 are applicable but they all lead to the same middle point with $x_9 = 0, y_9 = 0$. This means that the position of quadrature points is also independent of the center parameter.

```
element= 1, npts= 1000000, nqperdim= 3
========================================================
center= 0 x9,y9= 2.500000e-01 0.000000e+00
center= 0 min/max/avrg jcob on pts   = 3.250362e+00 4.249199e+00 3.750057e+00
center= 0 min/max/avrg err_f on pts  = -6.060741e+00 6.666524e+00 1.688878e-03
center= 0 min/max/avrg err_dfdx on pts= -1.099651e+01 6.443612e+00 -5.010148e-03
center= 0 min/max/avrg err_dfdy on pts= -1.015984e+01 4.225070e+00 1.686068e-02
center= 0 int err_f= 1.107770e+01
center= 0 int err_gradf= 1.954635e+01
center= 0 area= 1.500000e+01
========================================================
center= 1 x9,y9= 2.500000e-01 0.000000e+00
center= 1 min/max/avrg jcob on pts   = 3.250930e+00 4.249578e+00 3.749771e+00
center= 1 min/max/avrg err_f on pts  = -6.060172e+00 6.666509e+00 -4.038311e-03
center= 1 min/max/avrg err_dfdx on pts= -1.099842e+01 6.444102e+00 -1.172403e-04
center= 1 min/max/avrg err_dfdy on pts= -1.018956e+01 4.225064e+00 1.654201e-02
center= 1 int err_f= 1.107770e+01
center= 1 int err_gradf= 1.954635e+01
center= 1 area= 1.500000e+01
========================================================
center= 2 x9,y9= 2.500000e-01 0.000000e+00
center= 2 min/max/avrg jcob on pts   = 3.251054e+00 4.249622e+00 3.750130e+00
center= 2 min/max/avrg err_f on pts  = -6.060921e+00 6.666524e+00 -6.843429e-03
center= 2 min/max/avrg err_dfdx on pts= -1.099522e+01 6.441666e+00 -3.864851e-03
center= 2 min/max/avrg err_dfdy on pts= -1.018380e+01 4.225069e+00 1.319091e-02
center= 2 int err_f= 1.107770e+01
center= 2 int err_gradf= 1.954635e+01
center= 2 area= 1.500000e+01
========================================================
center= 6 x9,y9= 2.500000e-01 0.000000e+00
center= 6 min/max/avrg jcob on pts   = 3.250841e+00 4.249203e+00 3.749670e+00
center= 6 min/max/avrg err_f on pts  = -6.060064e+00 6.666475e+00 -9.101238e-04
center= 6 min/max/avrg err_dfdx on pts= -1.099351e+01 6.443416e+00 -7.271667e-03
center= 6 min/max/avrg err_dfdy on pts= -1.016288e+01 4.225062e+00 1.436911e-02
center= 6 int err_f= 1.107770e+01
center= 6 int err_gradf= 1.954635e+01
center= 6 area= 1.500000e+01
========================================================
```
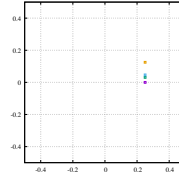
Conclusions:

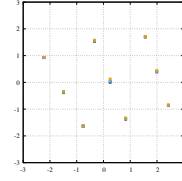- Exactly the same conclusions as for element 0 can be drawn.
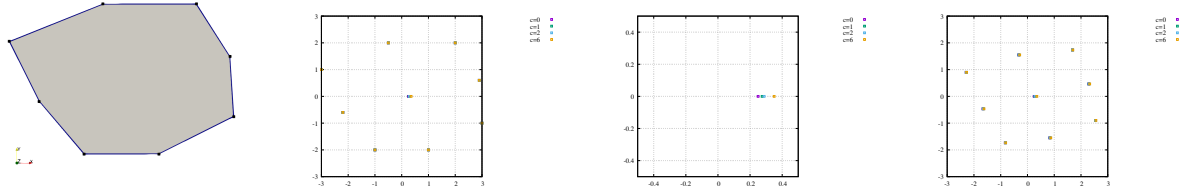
Element 2
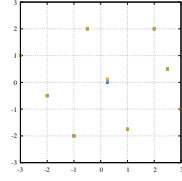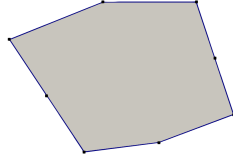

```
x1=-1  ;  y1=-2
x2=3   ;  y2=-1
x3=2   ;  y3=2
x4=-3  ;  y4=1
x5=0.5*(x1+x2)  ;  y5=0.5*(y1+y2)  -0.25
x6=0.5*(x2+x3)  ;  y6=0.5*(y2+y3)
x7=0.5*(x3+x4)  ;  y7=0.5*(y3+y4)  +0.5
x8=0.5*(x1+x4)  ;  y8=0.5*(y1+y4)
```

```
element= 2, npts= 1000000, nqperdim= 3
=========================================================
center= 0 x9,y9= 2.500000e-01 0.000000e+00
center= 0 min/max/avrg jcob on pts  = 3.003425e+00 5.833227e+00 4.332919e+00
center= 0 min/max/avrg err_f on pts   = -5.308564e+00 7.373735e+00 1.289018e-01
center= 0 min/max/avrg err_dfdx on pts= -1.323435e+01 6.260039e+00 -3.403394e-02
center= 0 min/max/avrg err_dfdy on pts= -8.721839e+00 5.063460e+00 -6.697600e-02
center= 0 int err_f= 1.183453e+01
center= 0 int err_gradf= 2.148351e+01
center= 0 area= 1.733333e+01
=========================================================
center= 1 x9,y9= 2.500000e-01 3.125000e-02
center= 1 min/max/avrg jcob on pts  = 3.002755e+00 5.696725e+00 4.333505e+00
center= 1 min/max/avrg err_f on pts   = -5.363672e+00 7.293081e+00 9.018116e-02
center= 1 min/max/avrg err_dfdx on pts= -1.323652e+01 6.221262e+00 -2.293697e-02
center= 1 min/max/avrg err_dfdy on pts= -8.682711e+00 5.047551e+00 -6.391543e-02
center= 1 int err_f= 1.178018e+01
center= 1 int err_gradf= 2.141089e+01
center= 1 area= 1.733333e+01
=========================================================
center= 2 x9,y9= 2.500000e-01 4.687500e-02
center= 2 min/max/avrg jcob on pts  = 3.002800e+00 5.630556e+00 4.332764e+00
center= 2 min/max/avrg err_f on pts   = -5.390923e+00 7.252336e+00 7.759085e-02
center= 2 min/max/avrg err_dfdx on pts= -1.321750e+01 6.199422e+00 -2.474020e-02
center= 2 min/max/avrg err_dfdy on pts= -8.715040e+00 5.038157e+00 -6.434275e-02
center= 2 int err_f= 1.175495e+01
center= 2 int err_gradf= 2.137246e+01
center= 2 area= 1.733333e+01
=========================================================
center= 6 x9,y9= 2.500000e-01 1.250000e-01
center= 6 min/max/avrg jcob on pts  = 3.003258e+00 5.340770e+00 4.332317e+00
center= 6 min/max/avrg err_f on pts   = -5.523826e+00 7.044686e+00 1.150397e-02
center= 6 min/max/avrg err_dfdx on pts= -1.321437e+01 6.083445e+00 -9.511932e-03
center= 6 min/max/avrg err_dfdy on pts= -8.679190e+00 4.976683e+00 -5.669887e-02
center= 6 int err_f= 1.164859e+01
center= 6 int err_gradf= 2.115913e+01
center= 6 area= 1.733333e+01
=========================================================
```

Conclusions:

- the four approaches yield 4 different locations for the center.

- E_f very comparable across all four cases. center=6 seems to be the best

- Same conclusion $E_{\nabla f}$

Element 3


```
x1=−1  ;  y1=−2
x2=3   ;  y2=−1
x3=2   ;  y3=2
x4=−3  ;  y4=1
x5=0.5*(x1+x2)      ;  y5=0.5*(y1+y2)  −0.5
x6=0.5*(x2+x3)+0.4  ;  y6=0.5*(y2+y3)  +0.1
x7=0.5*(x3+x4)      ;  y7=0.5*(y3+y4)  +0.5
x8=0.5*(x1+x4)−0.2  ;  y8=0.5*(y1+y4)  −0.1
```

```
element= 3, npts= 1000000, nqperdim= 3
============================================================
center= 0 x9,y9= 2.500000e-01 0.000000e+00
center= 0 min/max/avrg jcob on pts   = 1.907940e+00 5.649991e+00 4.849153e+00
center= 0 min/max/avrg err_f on pts   = -7.275654e+00 9.566299e+00 1.918971e-01
center= 0 min/max/avrg err_dfdx on pts= -1.653858e+01 6.294598e+00 -2.527334e-01
center= 0 min/max/avrg err_dfdy on pts= -2.581218e+01 5.947161e+00 -3.542729e-01
center= 0 int err_f= 1.614533e+01
center= 0 int err_gradf= 2.925334e+01
center= 0 area= 1.940000e+01
============================================================
center= 1 x9,y9= 2.750000e-01 1.387779e-17
center= 1 min/max/avrg jcob on pts   = 1.903598e+00 5.677286e+00 4.850195e+00
center= 1 min/max/avrg err_f on pts   = -7.297901e+00 9.493156e+00 1.465801e-01
center= 1 min/max/avrg err_dfdx on pts= -1.657761e+01 6.261874e+00 -2.442038e-01
center= 1 min/max/avrg err_dfdy on pts= -2.587333e+01 5.949630e+00 -3.426886e-01
center= 1 int err_f= 1.605547e+01
center= 1 int err_gradf= 2.915732e+01
center= 1 area= 1.940000e+01
============================================================
center= 2 x9,y9= 2.875000e-01 0.000000e+00
center= 2 min/max/avrg jcob on pts   = 1.904405e+00 5.693478e+00 4.849918e+00
center= 2 min/max/avrg err_f on pts   = -7.308703e+00 9.456916e+00 1.279976e-01
center= 2 min/max/avrg err_dfdx on pts= -1.668960e+01 6.246507e+00 -2.416973e-01
center= 2 min/max/avrg err_dfdy on pts= -2.604266e+01 5.951170e+00 -3.517167e-01
center= 2 int err_f= 1.601397e+01
center= 2 int err_gradf= 2.910870e+01
center= 2 area= 1.940000e+01
============================================================
center= 6 x9,y9= 3.500000e-01 5.551115e-17
center= 6 min/max/avrg jcob on pts   = 1.913670e+00 5.787307e+00 4.849397e+00
center= 6 min/max/avrg err_f on pts   = -7.359196e+00 9.279116e+00 1.625764e-02
center= 6 min/max/avrg err_dfdx on pts= -1.661175e+01 6.169925e+00 -2.370181e-01
center= 6 min/max/avrg err_dfdy on pts= -2.588039e+01 5.961927e+00 -3.386132e-01
center= 6 int err_f= 1.584097e+01
center= 6 int err_gradf= 2.885993e+01
center= 6 area= 1.940000e+01
============================================================
```
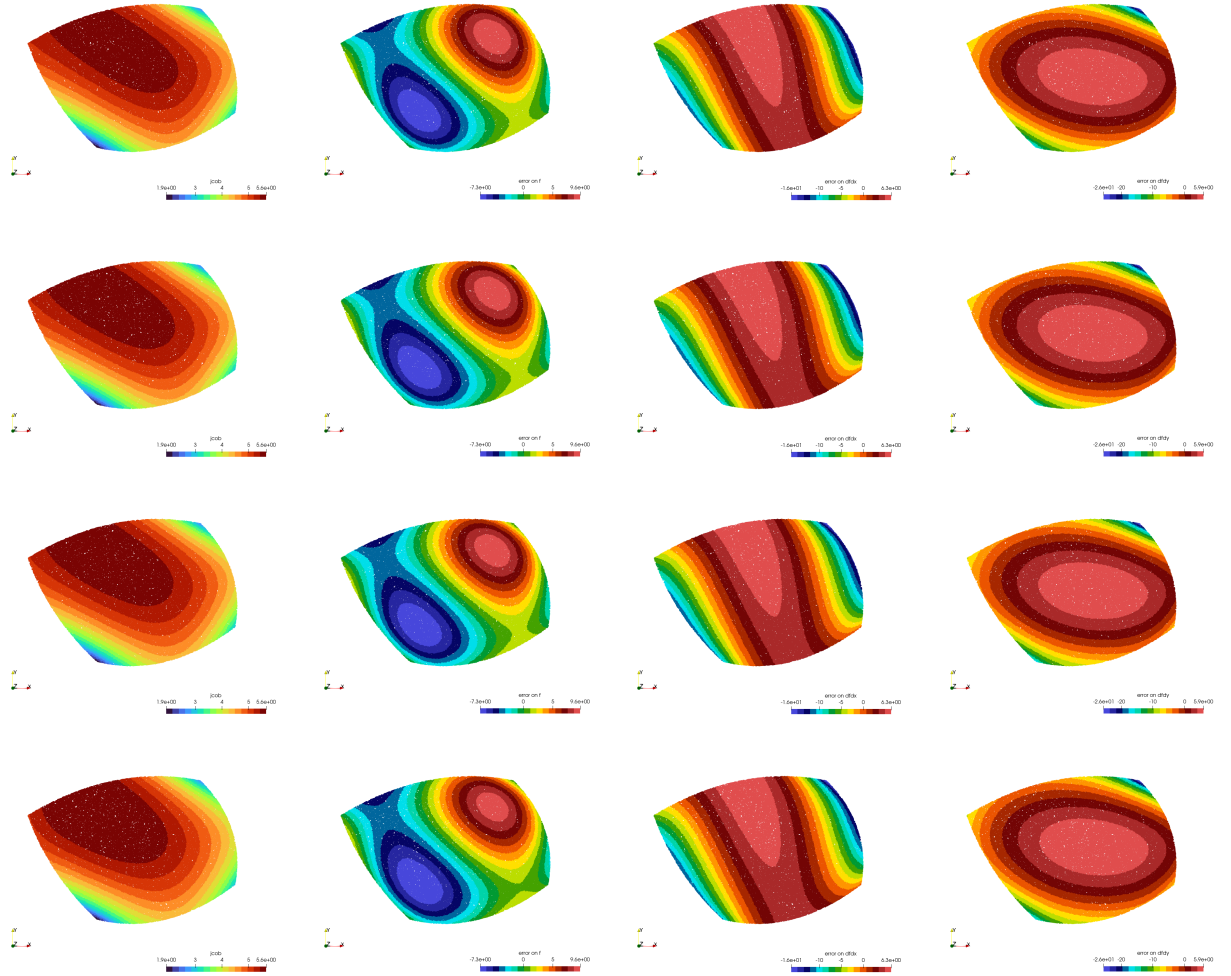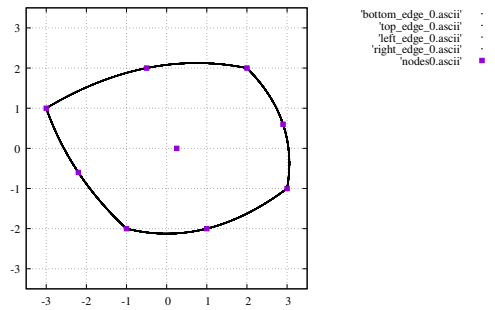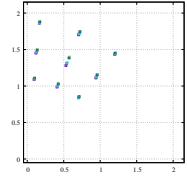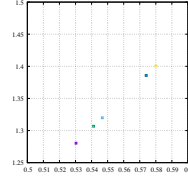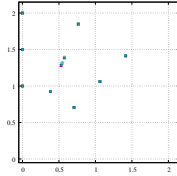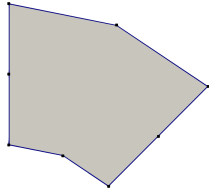
100,000 points at random. From left to right: jacobian, error on f, error on df/dx and error on df/dy. From top to bottom: center=0,1,2,6.

Conclusions:

- the four approaches yield 4 different locations for the center.

- E_f very comparable across all four cases. center=6 seems to be the best

- Same conclusion $E_{\nabla f}$

edges obtained by forward Q_2 mapping of points $r, s = \pm 1$

Element 4


```
R1=1
R2=2
dtheta=np.pi/4
x1=R1*np.cos(np.pi/2-0)                  ; y1=R1*np.sin(np.pi/2-0)
x5=R1*np.cos(np.pi/2-dtheta/2)           ; y5=R1*np.sin(np.pi/2-dtheta/2)
x2=R1*np.cos(np.pi/2-dtheta)             ; y2=R1*np.sin(np.pi/2-dtheta)
x8=0.5*(R1+R2)*np.cos(np.pi/2-0)         ; y8=0.5*(R1+R2)*np.sin(np.pi/2-0)
x6=0.5*(R1+R2)*np.cos(np.pi/2-dtheta)    ; y6=0.5*(R1+R2)*np.sin(np.pi/2-dtheta)
x4=R2*np.cos(np.pi/2-0)                  ; y4=R2*np.sin(np.pi/2-0)
x7=R2*np.cos(np.pi/2-dtheta/2)           ; y7=R2*np.sin(np.pi/2-dtheta/2)
x3=R2*np.cos(np.pi/2-dtheta)             ; y3=R2*np.sin(np.pi/2-dtheta)
```

```
center of mass coords (x,y): 0.5801028000103104 1.4004920473554983
center of mass coords (r,theta) 1.5158816686291174 67.5
element= 4
npts= 1000000
nqperdim= 3
========================================================
center= 0 x9,y9= 5.303301e-01 1.280330e+00
center= 0 min/max/avrg jcob on pts   = 1.039732e-01 5.574373e-01 2.943883e-01
center= 0 min/max/avrg err_f on pts   = -2.262961e-01 3.113327e-01 9.571182e-03
center= 0 min/max/avrg err_dfdx on pts= -1.424648e+00 6.317490e-01 7.838307e-03
center= 0 min/max/avrg err_dfdy on pts= -2.934500e+00 1.004204e+00 -2.939049e-02
center= 0 int err_f= 1.333177e-01
center= 0 int err_gradf= 9.290509e-01
center= 0 area= 1.177180e+00
========================================================
center= 1 x9,y9= 5.412539e-01 1.306702e+00
center= 1 min/max/avrg jcob on pts   = 1.258064e-01 5.136716e-01 2.941058e-01
center= 1 min/max/avrg err_f on pts   = -2.260227e-01 2.814548e-01 5.370194e-03
center= 1 min/max/avrg err_dfdx on pts= -1.425092e+00 6.022714e-01 1.754560e-02
center= 1 min/max/avrg err_dfdy on pts= -2.052857e+00 9.324228e-01 -1.974217e-02
center= 1 int err_f= 1.151357e-01
center= 1 int err_gradf= 8.033761e-01
center= 1 area= 1.177180e+00
========================================================
center= 2 x9,y9= 5.467157e-01 1.319889e+00
center= 2 min/max/avrg jcob on pts   = 1.367421e-01 4.919188e-01 2.944290e-01
center= 2 min/max/avrg err_f on pts   = -2.261415e-01 2.680915e-01 3.934531e-03
center= 2 min/max/avrg err_dfdx on pts= -1.426224e+00 5.884028e-01 2.200854e-02
center= 2 min/max/avrg err_dfdy on pts= -1.715733e+00 9.007872e-01 -1.831880e-02
center= 2 int err_f= 1.074743e-01
center= 2 int err_gradf= 7.456708e-01
center= 2 area= 1.177180e+00
========================================================
center= 3 x9,y9= 5.740251e-01 1.385819e+00
center= 3 min/max/avrg jcob on pts   = 1.913594e-01 4.118053e-01 2.942541e-01
center= 3 min/max/avrg err_f on pts   = -2.264001e-01 2.225621e-01 -6.321315e-04
center= 3 min/max/avrg err_dfdx on pts= -1.428531e+00 5.618227e-01 4.015243e-02
center= 3 min/max/avrg err_dfdy on pts= -8.218700e-01 7.798724e-01 -1.818832e-02
center= 3 int err_f= 8.691772e-02
center= 3 int err_gradf= 5.435321e-01
center= 3 area= 1.177180e+00
========================================================
center= 4 x9,y9= 5.801028e-01 1.400492e+00
center= 4 min/max/avrg jcob on pts   = 2.035020e-01 4.116166e-01 2.943053e-01
center= 4 min/max/avrg err_f on pts   = -2.263658e-01 2.219307e-01 -7.665193e-04
center= 4 min/max/avrg err_dfdx on pts= -1.431221e+00 6.504554e-01 4.500017e-02
center= 4 min/max/avrg err_dfdy on pts= -6.057840e-01 7.625320e-01 -1.938310e-02
center= 4 int err_f= 8.678766e-02
center= 4 int err_gradf= 5.294221e-01
center= 4 area= 1.177180e+00
========================================================
center= 6 x9,y9= 5.740251e-01 1.385819e+00
center= 6 min/max/avrg jcob on pts   = 1.913468e-01 4.117265e-01 2.942638e-01
center= 6 min/max/avrg err_f on pts   = -2.263541e-01 2.225620e-01 -5.845829e-04
center= 6 min/max/avrg err_dfdx on pts= -1.431369e+00 5.618216e-01 4.034998e-02
center= 6 min/max/avrg err_dfdy on pts= -8.219987e-01 7.798100e-01 -1.843770e-02
center= 6 int err_f= 8.691772e-02
center= 6 int err_gradf= 5.435321e-01
center= 6 area= 1.177180e+00
========================================================
```
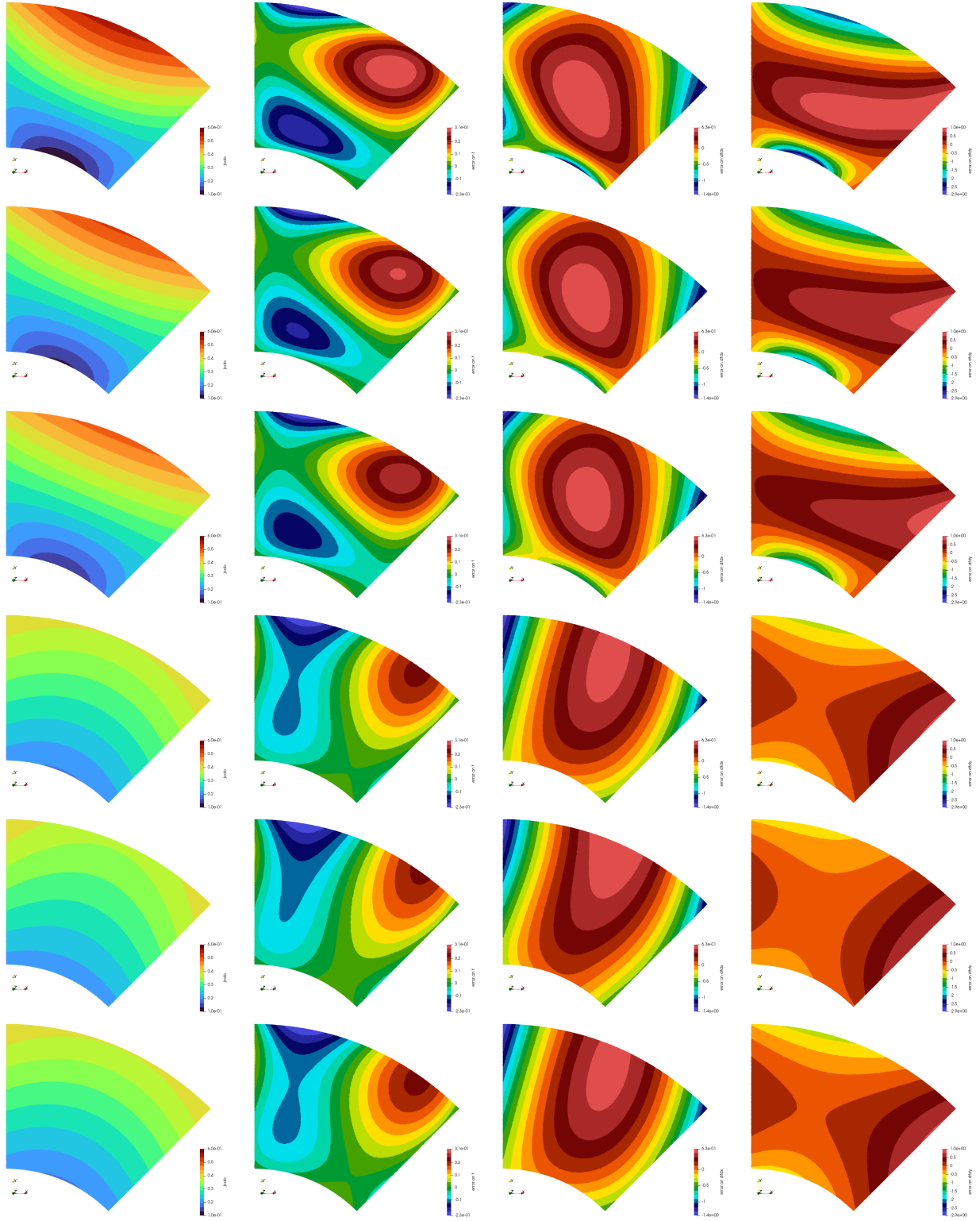
1,000,000 points at random. From left to right: jacobian, error on f, error on df/dx and error on df/dy. From top to bottom: center=0,1,2,3,4,6.

Conclusions:

- approaches 1,2,3 and 3,4,6 yield similar results respectively.

- fields also visually look different

- approaches 3,4,6 yield better results than 1,2,3.

- area still independent of approach

556

Relevant Literature

- K.Y. Yuan, Y.S. Huang, H.T. Yang, and T.H.H. Pian. "The inverse mapping and distortion measures for 8-node hexahedral isoparametric elements". In: *Computational Mechanics* 14 (1994), pp. 189–199. DOI: 10.1007/BF00350284

11.10 The Double Jacobian approach

What follows is very much inspired by Morgan, Taramón, and Hasenclever [959] (2020) with slight changes in the notations.

The basic idea behind this approach is to compute the local to Cartesian mapping as a two-stage process, hence the name "Double Jacobian".

1. The first stage maps from local to polar/spherical coordinates and back. This mapping is typically to a straight-edged polar or spherical element for which the Jacobian partial derivatives are constant within the element. The mapping and its inverse are given by straightforward analytical matrix expressions.

2. The second stage maps from polar/spherical to Cartesian coordinates (and back) and is also a simple analytical mapping.

The net Jacobian from local finite element coordinates to a cylindrical or spherical element in Cartesian geometry (or any other analytically mapped geometry) is simply the matrix product of two easy-to-compute inverse Jacobian matrices. Because the net Jacobian has an analytical form, it can be more rapidly computed than a general isoparametric or superparametric finite element mapping.

Remark: in what follows I will denote the polar coordinates by ρ, θ so as to avoid confusion with the reduced/local coordinates r, s.

```
                J_LC
            --------->
local (r,s)             Cartesian (x,z)
            <---------
                J_CL
```
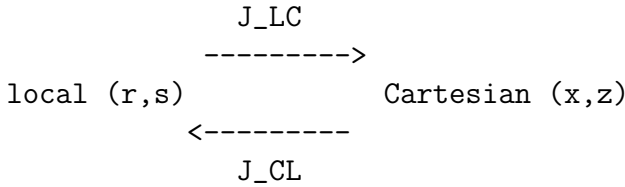
As we have seen so far the standard approach is as follows: The relationship between the derivatives of the basis functions with respect to the Cartesian coordinates and the same derivatives with respect to the local coordinates is given by:

$$
\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial s}{\partial x} \\[2mm] \frac{\partial r}{\partial z} & \frac{\partial s}{\partial z} \end{pmatrix}}_{\boldsymbol{J}_{LC}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix}
$$

where $\mathcal{N}_i$ are the shape functions, i is the local node numbering of the element, (x, z) are the Cartesian coordinates, and (r, s) are the local coordinates within the reference element. We of course have

$$
\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial z}{\partial r} \\[2mm] \frac{\partial x}{\partial s} & \frac{\partial z}{\partial s} \end{pmatrix}}_{\boldsymbol{J}_{CL}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix}
$$

Using an isoparametric mapping:

$$
\boldsymbol{J}_{CL} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial z}{\partial r} \\[2mm] \frac{\partial x}{\partial s} & \frac{\partial z}{\partial s} \end{pmatrix} = \begin{pmatrix} \frac{\partial (\sum_i \mathcal{N}_i x_i)}{\partial r} & \frac{\partial (\sum_i \mathcal{N}_i z_i)}{\partial r} \\[2mm] \frac{\partial (\sum_i \mathcal{N}_i x_i)}{\partial s} & \frac{\partial (\sum_i \mathcal{N}_i z_i)}{\partial s} \end{pmatrix} = \begin{pmatrix} \sum_i \frac{\partial \mathcal{N}_i}{\partial r} x_i & \sum_i \frac{\partial \mathcal{N}_i}{\partial r} z_i \\[2mm] \sum_i \frac{\partial \mathcal{N}_i}{\partial s} x_i & \sum_i \frac{\partial \mathcal{N}_i}{\partial s} z_i \end{pmatrix}
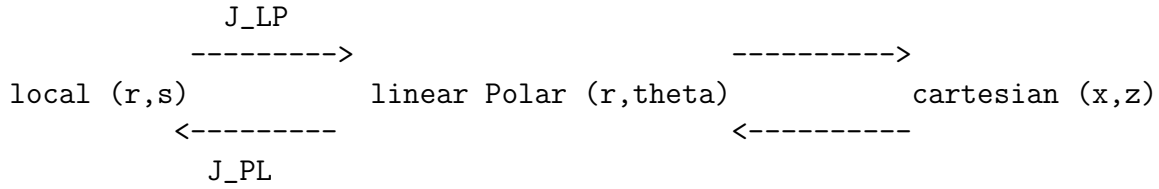$$

In the end

$$
\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \boldsymbol{J}_{CL}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} = \begin{pmatrix} \sum_i \frac{\partial \mathcal{N}_i}{\partial r} x_i & \sum_i \frac{\partial \mathcal{N}_i}{\partial r} z_i \\[2mm] \sum_i \frac{\partial \mathcal{N}_i}{\partial s} x_i & \sum_i \frac{\partial \mathcal{N}_i}{\partial s} z_i \end{pmatrix}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix}
\tag{11.20}
$$

In many stones this is coded as follows:

```
jcb=np.zeros((2,2),dtype=np.float64)
for k in range(0,mV):
    jcb[0,0] += dNNNVdr[k]*xV[iconV[k,iel]]
    jcb[0,1] += dNNNVdr[k]*yV[iconV[k,iel]]
    jcb[1,0] += dNNNVds[k]*xV[iconV[k,iel]]
    jcb[1,1] += dNNNVds[k]*yV[iconV[k,iel]]
jcob = np.linalg.det(jcb)
jcbi = np.linalg.inv(jcb)
for k in range(0,mV):
    dNNNVdx[k]=jcbi[0,0]*dNNNVdr[k]+jcbi[0,1]*dNNNVds[k]
    dNNNVdy[k]=jcbi[1,0]*dNNNVdr[k]+jcbi[1,1]*dNNNVds[k]
```

11.10.1 First Jacobian

```
                    J_LP
           --------->                 ---------->
 local (r,s)            linear Polar (r,theta)            cartesian (x,z)
           <---------                 <----------
           J_PL
```

The Double Jacobian approach uses the standard finite element approach to first map from local to linear polar coordinates (i.e. ρ, θ). The first Jacobian for a cylindrical (polar) mapping is analogous to the standard Jacobian, where x and z are now changed to ρ and θ, respectively. The global derivatives may be expressed in matrix form as

$$
\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial \rho} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial \theta} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial r}{\partial \rho} & \frac{\partial s}{\partial \rho} \\[2mm] \frac{\partial r}{\partial \theta} & \frac{\partial s}{\partial \theta} \end{pmatrix}}_{\boldsymbol{J}_{LP}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix}
\tag{11.21}
$$

where $\mathcal{N}_i$ are the shape functions and $\boldsymbol{J}_{LP}$ is the Jacobian of the transformation from local to polar coordinates.

The derivatives of these shape functions with respect to local coordinates r, s (i.e. the rhs vector) can be computed explicitly since the basis functions are chosen/built for a given element type and formulated as a function of the local coordinates. However, the terms of the Jacobian $\boldsymbol{J}_{LP}$ cannot be directly computed since explicit expressions for $r(\rho, \theta)$ and $s(\rho, \theta)$ do not exist. A wonderful "trick" in finite element programming (discovered by Bruce Irons in the mid-60s [402]) is to make use of the inverse coordinate transformation

$$
\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial \rho}{\partial r} & \frac{\partial \theta}{\partial r} \\[2mm] \frac{\partial \rho}{\partial s} & \frac{\partial \theta}{\partial s} \end{pmatrix}}_{\boldsymbol{J}_{PL}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial \rho} \\[2mm] \frac{\partial \mathcal{N}_i}{\partial \theta} \end{pmatrix}
\tag{11.22}
$$

where $i\boldsymbol{J}_{PL}$ is the Jacobian of the transformation from polar to local coordinates.

From Equations (11.21) and (11.22), we have $\boldsymbol{J}_{LP} = \boldsymbol{J}_{PL}^{-1}$. Using Cramer's rule, the inverse of the Jacobian from polar to local coordinates is given by

$$\boldsymbol{J}_{LP} = \boldsymbol{J}_{PL}^{-1} = \frac{1}{|\boldsymbol{J}_{PL}|} \begin{pmatrix} \frac{\partial \theta}{\partial s} & -\frac{\partial \theta}{\partial r} \\ -\frac{\partial \rho}{\partial s} & \frac{\partial \rho}{\partial r} \end{pmatrix}$$

The polar coordinates for each element are related with local coordinates through the shape functions:

$$\rho(r, s) = \sum_{i=1}^{m} \mathcal{N}_i(r, s)\, \rho_i$$

$$\theta(r, s) = \sum_{i=1}^{m} \mathcal{N}_i(r, s)\, \theta_i$$

where m is the number of nodes in the element. Then

$$\frac{\partial \rho}{\partial r}(r, s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r, s)\, \rho_i$$

$$\frac{\partial \rho}{\partial s}(r, s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r, s)\, \rho_i$$

$$\frac{\partial \theta}{\partial r}(r, s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r, s)\, \theta_i$$

$$\frac{\partial \theta}{\partial s}(r, s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r, s)\, \theta_i \tag{11.23}$$

11.10.2   Second jacobian

The second Jacobian in the DJ method is the *analytical* mapping from polar coordinates to Cartesian coordinates. The derivatives expressed in matrix form are given by

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial \rho}{\partial x} & \frac{\partial \theta}{\partial x} \\ \frac{\partial \rho}{\partial z} & \frac{\partial \theta}{\partial z} \end{pmatrix}}_{\boldsymbol{J}_{PC}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial \rho} \\ \frac{\partial \mathcal{N}_i}{\partial \theta} \end{pmatrix} \tag{11.24}$$

where $\boldsymbol{J}_{PC}$ is the Jacobian from polar to Cartesian coordinates. The analytical expressions for $\theta(x, z)$ and $\rho(x, z)$ are known; however, it is again easier to use the inverse transformation

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial \rho} \\ \frac{\partial \mathcal{N}_i}{\partial \theta} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial z}{\partial \rho} \\ \frac{\partial x}{\partial \theta} & \frac{\partial z}{\partial \theta} \end{pmatrix}}_{\boldsymbol{J}_{CP}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial y} \end{pmatrix} \tag{11.25}$$

where $\boldsymbol{J}_{CP}$ is the Jacobian from Cartesian to polar coordinates. From Equations (11.24) and (11.25), $\boldsymbol{J}_{PC} = \boldsymbol{J}_{CP}^{-1}$. The inverse of the Jacobian from Cartesian to polar coordinates is given by

$$\boldsymbol{J}_{PC} = \boldsymbol{J}_{CP}^{-1} = \frac{1}{|\boldsymbol{J}_{CP}|} \begin{pmatrix} \frac{\partial z}{\partial \theta} & -\frac{\partial z}{\partial \rho} \\ -\frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \rho} \end{pmatrix}$$

Cartesian coordinates are related with polar coordinates by:

$$x(\rho, \theta) = \rho \cos \theta$$
$$z(\rho, \theta) = \rho \sin \theta$$

so that

$$\frac{\partial x}{\partial \rho}(\theta, \rho) = \cos \theta \tag{11.26}$$

$$\frac{\partial x}{\partial \theta}(\theta, \rho) = -\rho \sin \theta \tag{11.27}$$

$$\frac{\partial z}{\partial \rho}(\theta, \rho) = \sin \theta \tag{11.28}$$

$$\frac{\partial z}{\partial \theta}(\theta, \rho) = \rho \cos \theta \tag{11.29}$$

Then

$$\boldsymbol{J}_{CP} = \begin{pmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial z}{\partial \rho} \\ \frac{\partial x}{\partial \theta} & \frac{\partial z}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\rho \sin \theta & \rho \cos \theta \end{pmatrix}$$

with obviously $det(\boldsymbol{J}_{CP}) = \rho \cos^2 \theta + \rho \sin^2 \theta = \rho$.

The inverse of the Jacobian from Cartesian coordinates to polar coordinates can then be written as a function of polar coordinates:

$$\boxed{\boldsymbol{J}_{CP}^{-1} = \frac{1}{\rho} \begin{pmatrix} \rho \cos \theta & -\sin \theta \\ \rho \sin \theta & \cos \theta \end{pmatrix}} \tag{11.30}$$

where θ and ρ are evaluated at each integration point (ip).

Remark: this Jacobian is independent of the choice of basis functions since it is analytical.

11.10.3 Combining both

Making use of the matrix product of the two inverse Jacobians, global Cartesian derivatives can be expressed as a matrix product of the local derivatives in the local to polar and polar to Cartesian coordinate mappings. Substituting Equation (11.21) into Equation (11.24) yields

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \boldsymbol{J}_{PC} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial \rho} \\ \frac{\partial \mathcal{N}_i}{\partial \theta} \end{pmatrix} = \boldsymbol{J}_{PC} \cdot \boldsymbol{J}_{LP} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} = \boldsymbol{J}_{CP}^{-1} \cdot \boldsymbol{J}_{PL}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} \tag{11.31}$$

Note that the 2d DJ approach will ensure a perfect mapping to the circular-arc edges of the elements of a cylindrical annulus mesh. Any point on the edge of a boundary element of the mesh is mapped to its true position along a circular arc.

The major drawback from this is the fact that (as we will see) the Jacobian is no more a polynomial so special care must be taken with regards to the integration.

11.10.4 Linear triangles (P_1)

I need to revisit this.. this is still with rho and theta as in the paper.

For linear triangles (P_1) the basis functions are (see Section 8.3.10):

$$\mathcal{N}_1(r, s) = 1 - r - s$$
$$\mathcal{N}_2(r, s) = r$$
$$\mathcal{N}_3(r, s) = s$$

and their derivatives:

$$\frac{\partial \theta}{\partial r}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\, \theta_i = \theta_2 - \theta_1 = \theta_{21}$$

$$\frac{\partial \theta}{\partial s}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\, \theta_i = \theta_3 - \theta_1 = \theta_{31}$$

$$\frac{\partial \rho}{\partial r}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\, \rho_i = \rho_2 - \rho_1 = \rho_{21}$$

$$\frac{\partial \rho}{\partial s}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\, \rho_i = \rho_3 - \rho_1 = \rho_{31}$$

In this case we have

$$\boxed{\boldsymbol{J}_{LP} = \boldsymbol{J}_{PL}^{-1} = \frac{1}{\theta_{21}\rho_{31} - \rho_{21}\theta_{31}} \begin{pmatrix} \rho_{31} & -\rho_{21} \\ -\theta_{31} & \theta_{21} \end{pmatrix}} \tag{11.32}$$

Substituting Equations (11.32) and (11.30) into Eq. (11.31), the analytical expression for this mapping is

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \frac{1}{\rho} \begin{pmatrix} \cos\theta & \rho\sin\theta \\ -\sin\theta & \rho\cos\theta \end{pmatrix} \cdot \frac{1}{\theta_{21}\rho_{31} - \rho_{21}\theta_{31}} \begin{pmatrix} \rho_{31} & -\rho_{21} \\ -\theta_{31} & \theta_{21} \end{pmatrix}$$

$$= \frac{1}{\theta_{21}\rho_{31} - \rho_{21}\theta_{31}} \begin{pmatrix} \frac{\rho_{31}}{\rho}\cos\theta - \theta_{31}\sin\theta & -\frac{\rho_{21}}{\rho}\cos\theta + \theta_{21}\sin\theta \\ -\frac{\rho_{31}}{\rho}\sin\theta - \theta_{31}\cos\theta & \frac{\rho_{21}}{\rho}\sin\theta + \theta_{21}\cos\theta \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} \tag{11.33}$$

which is Eq. (19) of Morgan, Taramón, and Hasenclever [959]. This expression could be problematic if $\rho = 0$ but since it will be evaluated at the quadrature points this case is extremely unlikely.

Based on Fig 2 of Morgan, Taramón, and Hasenclever [959], we consider an annulus of outer diameter 6371km. Visually from 2A we infer $\rho_1 = 4700$ km, $\rho_2 = \rho_3 = 6371$ km, and $\theta_1 = 14.5°$, $\theta_2 = 22°$ and $\theta_3 = 0°$.

If quadratic basis functions are used, we define $r_4 = \frac{1}{2}(r_1 + r_2)$, $r_5 = \frac{1}{2}(r_2 + r_3)$, $r_6 = \frac{1}{2}(r_3 + r_1)$, and likewise for the $\theta_{4,5,6}$ values.

We then proceed to generate 1000 random points in the reference triangle and plot their image in the Cartesian plane, either using the DJ method, a linear mapping P_1 or a quadratic mapping P_2. The python code is in `images/mappings/DJ/triangles`.

Unsurprisingly the P_1 mapping yields a triangle which 2-3 edge does not conform to the edge of the domain. The P_2 mapping does a much better job and so does the DJ mapping.
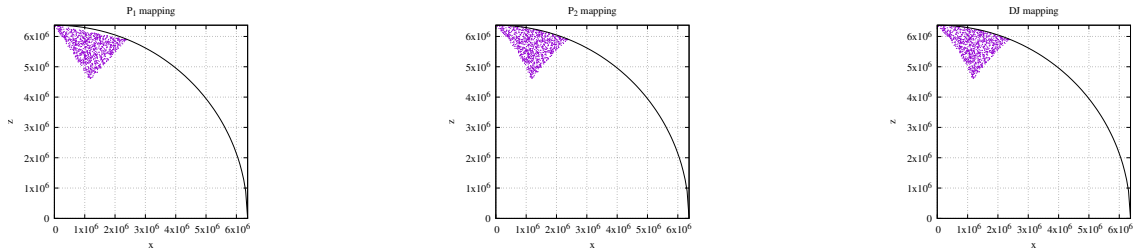
Let us now generate 1000 points on the 2-3 edge of the reference triangle and study their image with the three mappings:
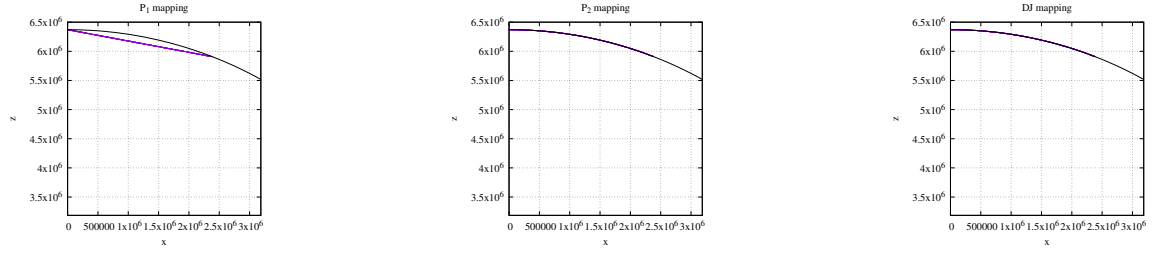
We cannot see any different between P2 and DJ. Let us now plot the error, i.e. the distance of the image point to the true surface $\rho = 6371$km as a function of x:
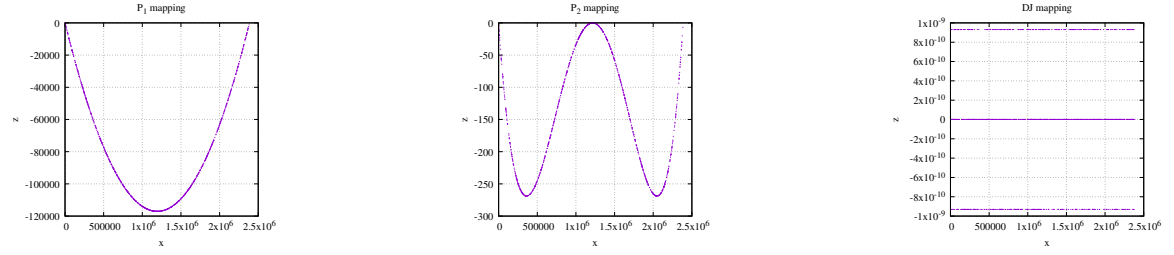
We find that the maximum error for P_1 is about 120km, the error for P_2 is about 260m, and the error for DJ is effectively about $< 10^{-9}$m.

11.10.5 Bilinear quadrilaterals (Q_1)

For linear quadrilaterals (Q_1) the basis functions are:

$$
\begin{aligned}
\mathcal{N}_1(r,s) &= \frac{1}{4}(1-r)(1-s) \\
\mathcal{N}_2(r,s) &= \frac{1}{4}(1+r)(1-s) \\
\mathcal{N}_3(r,s) &= \frac{1}{4}(1+r)(1+s) \\
\mathcal{N}_4(r,s) &= \frac{1}{4}(1-r)(1+s)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \rho}{\partial r}(r,s) &= \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\,\rho_i \\
&= -\frac{1}{4}(1-s)\rho_1 + \frac{1}{4}(1-s)\rho_2 + \frac{1}{4}(1+s)\rho_3 - \frac{1}{4}(1+s)\rho_4 \\
&= \frac{1}{4}(-\rho_1 + \rho_2 + \rho_3 - \rho_4) + \frac{1}{4}s(\rho_1 - \rho_2 + \rho_3 - \rho_4) \\
&= \tilde{\rho}_{1234} + \underline{\rho}_{1234}s \\
\frac{\partial \rho}{\partial s}(r,s) &= \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\rho_i \\
&= -\frac{1}{4}(1-r)\rho_1 - \frac{1}{4}(1+r)\rho_2 + \frac{1}{4}(1+r)\rho_3 + \frac{1}{4}(1-r)\rho_4 \\
&= \frac{1}{4}(-\rho_1 - \rho_2 + \rho_3 + \rho_4) + \frac{1}{4}r(\rho_1 - \rho_2 + \rho_3 - \rho_4) \\
&= \bar{\rho}_{1234} + \underline{\rho}_{1234}r \\
\frac{\partial \theta}{\partial r}(r,s) &= \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\,\theta_i \\
&= -\frac{1}{4}(1-s)\theta_1 + \frac{1}{4}(1-s)\theta_2 + \frac{1}{4}(1+s)\theta_3 - \frac{1}{4}(1+s)\theta_4 \\
&= \frac{1}{4}(-\theta_1 + \theta_2 + \theta_3 - \theta_4) + \frac{1}{4}s(\theta_1 - \theta_2 + \theta_3 - \theta_4) \\
&= \tilde{\theta}_{1234} + \underline{\theta}_{1234}s \\
\frac{\partial \theta}{\partial s}(r,s) &= \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\theta_i \\
&= -\frac{1}{4}(1-r)\theta_1 - \frac{1}{4}(1+r)\theta_2 + \frac{1}{4}(1+r)\theta_3 + \frac{1}{4}(1-r)\theta_4 \\
&= \frac{1}{4}(-\theta_1 - \theta_2 + \theta_3 + \theta_4) + \frac{1}{4}r(\theta_1 - \theta_2 + \theta_3 - \theta_4) \\
&= \bar{\theta}_{1234} + \underline{\theta}_{1234}r
\end{aligned}
$$

which leads to write

$$
\boldsymbol{J}_{PL} = \begin{pmatrix} \frac{\partial \rho}{\partial r} & \frac{\partial \theta}{\partial r} \\ \frac{\partial \rho}{\partial s} & \frac{\partial \theta}{\partial s} \end{pmatrix} = \begin{pmatrix} \tilde{\rho}_{1234} + \underline{\rho}_{1234}s & \tilde{\theta}_{1234} + \underline{\theta}_{1234}s \\ \bar{\rho}_{1234} + \underline{\rho}_{1234}r & \bar{\theta}_{1234} + \underline{\theta}_{1234}r \end{pmatrix}
$$

In this case we have $\boldsymbol{J}_{LP} = \boldsymbol{J}_{PL}^{-1}$ with

$$\boldsymbol{J}_{PL}^{-1} = \frac{1}{(\tilde{\rho}_{1234} + \underline{\rho}_{1234}s)(\bar{\theta}_{1234} + \underline{\theta}_{1234}r) - (\bar{\rho}_{1234} + \underline{\rho}_{1234}r)(\tilde{\theta}_{1234} + \underline{\theta}_{1234}s)} \begin{pmatrix} \bar{\theta}_{1234} + \underline{\theta}_{1234}r & -\tilde{\theta}_{1234} - \underline{\theta}_{1234}s \\ \\ -\bar{\rho}_{1234} - \underline{\rho}_{1234}r & \tilde{\rho}_{1234} + \underline{\rho}_{1234}s \end{pmatrix}$$
(11.34)

SPECIAL CASE: In the element we have $\theta_1 = \theta_2$, $\theta_3 = \theta_4$, and we define $\tilde{\theta}$ such that $\frac{1}{2}(\theta_4 - \theta_1) = \frac{1}{2}(\theta_3 - \theta_2) = \bar{\theta}$. Likewise, $\rho_1 = \rho_4$, $\rho_2 = \rho_3$ and we define $\bar{\rho}$ such that $\frac{1}{2}(\rho_2 - \rho_1) = \frac{1}{2}(\rho_3 - \rho_4) = \tilde{\rho}$

```
uptheta
|
|
|    4---------3
|    |         |
|    |         |
|    |         |
-----1--------2---------- uprho
```
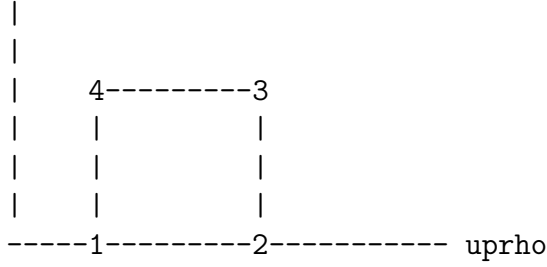
In this case:

$$\tilde{\rho}_{1234} = \frac{1}{4}(-\rho_1 + \rho_2 + \rho_3 - \rho_4) = \tilde{\rho}$$

$$\underline{\rho}_{1234} = \frac{1}{4}(\rho_1 - \rho_2 + \rho_3 - \rho_4) = 0$$

$$\bar{\rho}_{1234} = \frac{1}{4}(-\rho_1 - \rho_2 + \rho_3 + \rho_4) = 0$$

$$\tilde{\theta}_{1234} = \frac{1}{4}(-\theta_1 + \theta_2 + \theta_3 - \theta_4) = 0$$

$$\underline{\theta}_{1234} = \frac{1}{4}(\theta_1 - \theta_2 + \theta_3 - \theta_4) = 0$$

$$\bar{\theta}_{1234} = \frac{1}{4}(-\theta_1 - \theta_2 + \theta_3 + \theta_4) = \bar{\theta} \qquad (11.35)$$

so that

$$\boldsymbol{J}_{PL} = \begin{pmatrix} \frac{\partial \rho}{\partial r} & \frac{\partial \theta}{\partial r} \\ \frac{\partial \rho}{\partial s} & \frac{\partial \theta}{\partial s} \end{pmatrix} = \begin{pmatrix} \tilde{\rho} & 0 \\ 0 & \bar{\theta} \end{pmatrix} \qquad \Rightarrow \qquad \boldsymbol{J}_{LP} = \boldsymbol{J}_{PL}^{-1} = \begin{pmatrix} \tilde{\rho}^{-1} & 0 \\ 0 & \bar{\theta}^{-1} \end{pmatrix}$$

Bringing things together in Eq. (11.31):

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \boldsymbol{J}_{CP}^{-1} \cdot \boldsymbol{J}_{PL}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} = \frac{1}{\rho} \begin{pmatrix} \rho \cos\theta & -\sin\theta \\ \rho \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} \tilde{\rho}^{-1} & 0 \\ 0 & \bar{\theta}^{-1} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix}$$
(11.36)

$$\begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial x} \\ \frac{\partial \mathcal{N}_i}{\partial z} \end{pmatrix} = \underbrace{\begin{pmatrix} \tilde{\rho}^{-1} \cos\theta & -\bar{\theta}^{-1}\rho^{-1}\sin\theta \\ \tilde{\rho}^{-1}\sin\theta & \bar{\theta}^{-1}\rho^{-1}\cos\theta \end{pmatrix}}_{\boldsymbol{J}_{LC}} \cdot \begin{pmatrix} \frac{\partial \mathcal{N}_i}{\partial r} \\ \frac{\partial \mathcal{N}_i}{\partial s} \end{pmatrix} \qquad (11.37)$$

with

$$
\begin{aligned}
det(\boldsymbol{J}_{LC}) &= (\tilde{\rho}^{-1}\cos\theta)(\bar{\theta}^{-1}\rho^{-1}\cos\theta) + (\tilde{\rho}^{-1}\sin\theta)(\bar{\theta}^{-1}\rho^{-1}\sin\theta) \\
&= \rho^{-1}\left[\tilde{\rho}^{-1}\bar{\theta}^{-1}\cos^2\theta + \tilde{\rho}^{-1}\bar{\theta}^{-1}\sin^2\theta\right] \\
&= \rho^{-1}\tilde{\rho}^{-1}\bar{\theta}^{-1}\left[\cos^2\theta + \sin^2\theta\right] \\
&= \frac{1}{\rho\tilde{\rho}\bar{\theta}}
\end{aligned}
\tag{11.38}
$$

If matrix $\boldsymbol{A}$ is invertible then $\boldsymbol{A}\cdot\boldsymbol{A}^{-1} = \boldsymbol{1}$, and since $det(\boldsymbol{A}\cdot\boldsymbol{B}) = det(\boldsymbol{A})det(\boldsymbol{B})$, then $det(\boldsymbol{A}\cdot\boldsymbol{A}^{-1}) = det(\boldsymbol{A})\,det(\boldsymbol{A}^{-1}) = 1$. This means that

$$
det(\boldsymbol{J}_{CL}) = det(\boldsymbol{J}_{LC}^{-1}) = [det(\boldsymbol{J}_{LC})]^{-1} = \rho\tilde{\rho}\bar{\theta}
$$

Finally, from Cramer's rule[8], we obtain:

$$
\begin{aligned}
\boldsymbol{J}_{CL} = \boldsymbol{J}_{LC}^{-1} &= \frac{1}{det(\boldsymbol{J}_{LC})}\begin{pmatrix} \bar{\theta}^{-1}\rho^{-1}\cos\theta & \bar{\theta}^{-1}\rho^{-1}\sin\theta \\ -\tilde{\rho}^{-1}\sin\theta & \tilde{\rho}^{-1}\cos\theta \end{pmatrix} \\
&= \rho\tilde{\rho}\bar{\theta}\begin{pmatrix} \bar{\theta}^{-1}\rho^{-1}\cos\theta & \bar{\theta}^{-1}\rho^{-1}\sin\theta \\ -\tilde{\rho}^{-1}\sin\theta & \tilde{\rho}^{-1}\cos\theta \end{pmatrix} \\
&= \tilde{\rho}\bar{\theta}\begin{pmatrix} \bar{\theta}^{-1}\cos\theta & \bar{\theta}^{-1}\sin\theta \\ -\rho\tilde{\rho}^{-1}\sin\theta & \rho\tilde{\rho}^{-1}\cos\theta \end{pmatrix} \\
&= \begin{pmatrix} \tilde{\rho}\cos\theta & \tilde{\rho}\sin\theta \\ -\rho\bar{\theta}\sin\theta & \rho\bar{\theta}\cos\theta \end{pmatrix}
\end{aligned}
\tag{11.39}
$$

and of course we recover $det(\boldsymbol{J}_{CL}) = \rho\bar{\theta}\tilde{\rho}$. Finally:

$$
\boxed{\boldsymbol{J}_{CL} = \begin{pmatrix} \tilde{\rho}\cos\theta & \tilde{\rho}\sin\theta \\ -\rho\bar{\theta}\sin\theta & \rho\bar{\theta}\cos\theta \end{pmatrix}}
$$

Note that the area of such element can easily be computed analytically:

$$
\mathcal{A}_e = \iint_{\Omega_e}\rho\,d\rho\,d\theta = \int_{\rho_1}^{\rho_3}\int_{\theta_3}^{\theta_1}\rho\,d\rho\,d\theta = (\theta_1 - \theta_3)\frac{1}{2}(\rho_3^2 - \rho_1^2)
$$

We can now compute the same area but using a numerical quadrature:
(iq: in theta direction, jq in r direction) sum w= 2

$$
\begin{aligned}
\mathcal{A}_e &= \iint_{\Omega_e} dV \\
&= \sum_{iq=1}^{n_q}\sum_{jq=1}^{n_q} w_{iq}w_{jq}det(\boldsymbol{J})\big|_{i_q} \\
&= \sum_{iq=1}^{n_q}\sum_{jq=1}^{n_q} w_{iq}w_{jq}(\rho\bar{\rho}\tilde{\theta}) \\
&= \tilde{\theta}\bar{\rho}\sum_{iq=1}^{n_q}\sum_{jq=1}^{n_q} w_{iq}w_{jq}\rho_{iq,jq} \\
&= \tilde{\theta}\bar{\rho}\,2\sum_{jq=1}^{n_q} w_{jq}\rho_{jq}
\end{aligned}
\tag{11.40}
$$

---

[8]https://en.wikipedia.org/wiki/Cramers_rule

We have, since $\rho = mr + p$ with $p = \frac{1}{2}(\rho_1 + \rho_3)$ in the element (see Example 2 in Section 7.1.4):

$$\sum_{jq=1}^{n_q} w_{jq}\rho_{jq} = \frac{1}{2}(\rho_1 + \rho_3) \cdot 2 = \rho_1 + \rho_3$$

so that

$$
\begin{aligned}
\mathcal{A}_e &= \tilde{\theta}\bar{\rho}2(\rho_1 + \rho_3) \\
&= \frac{1}{2}(\theta_3 - \theta_1)\frac{1}{2}(\rho_3 - \rho_1)2(\rho_1 + \rho_3) \\
&= \frac{1}{2}(\theta_3 - \theta_1)(\rho_3^2 - \rho_1^2)
\end{aligned}
\tag{11.41}
$$

which is indeed the area of an element.

Let us consider an annulus with inner radius R_1 and outer radius R_2. Let us assume it is tesselated by means of non-overlapping elements, all characterised by the same $\delta\rho$ and $\delta\theta$ dimensions.

$$
\begin{aligned}
\mathcal{A} &= \iint_\Omega dV \\
&= \sum_e \iint_{\Omega_e} dV \\
&= \sum_e \mathcal{A}_e \\
&= \sum_e \sum_{iq=1}^{n_q} \sum_{jq=1}^{n_q} w_{iq}w_{jq}det(\boldsymbol{J})|_{i_q} \\
&= \sum_e \sum_{iq=1}^{n_q} \sum_{jq=1}^{n_q} w_{iq}w_{jq}(\rho\frac{1}{2}\delta\theta\frac{1}{2}\delta\rho) \\
&= \sum_e (\frac{1}{2}\delta\theta\frac{1}{2}\delta\rho)_e \sum_{iq=1}^{n_q} \sum_{jq=1}^{n_q} w_{iq}w_{jq}\rho_{iq,jq} \\
&= \sum_e (\frac{1}{2}\delta\theta\frac{1}{2}\delta\rho)_e 2 \sum_{jq=1}^{n_q} w_{jq}\rho_{jq}
\end{aligned}
\tag{11.42}
$$

FINISH/CORRECT

We can carry out the same exercise as before. The python code is in `images/mappings/DJ/quadrilateral`
The element is defined by $\theta_1 = \theta_4 = 10°$, $\theta_2 = \theta_3 = 25°$, $\rho_1 = \rho_2 = 4700$ km, $\rho_3 = \rho_4 = 6371$ km.
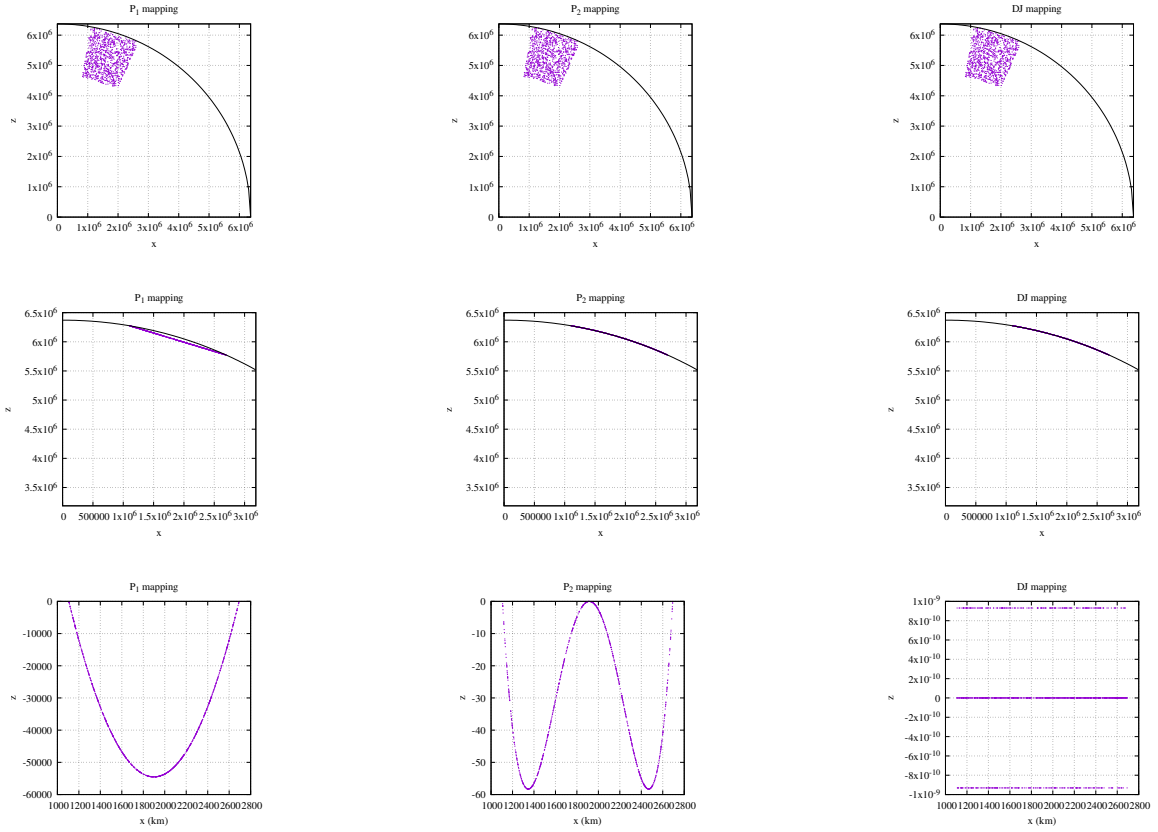
Quadratic quadrilaterals (Q_2)

The layout of the nodes is as follows:

```
4 7 3
8 9 6
1 5 2
```

The basis functions are given by

$$
\begin{aligned}
\mathcal{N}_1(r,s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \\
\mathcal{N}_2(r,s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \\
\mathcal{N}_3(r,s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s+1) \\
\mathcal{N}_4(r,s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \\
\mathcal{N}_5(r,s) &= (1-r^2)\frac{1}{2}s(s-1) \\
\mathcal{N}_6(r,s) &= \frac{1}{2}r(r+1)(1-s^2) \\
\mathcal{N}_7(r,s) &= (1-r^2)\frac{1}{2}s(s+1) \\
\mathcal{N}_8(r,s) &= \frac{1}{2}r(r-1)(1-s^2) \\
\mathcal{N}_9(r,s) &= (1-r^2)(1-s^2)
\end{aligned}
$$

568

with the following derivatives:

$$\frac{\partial \mathcal{N}_1}{\partial r} = \frac{1}{2}(2r-1)\frac{1}{2}s(s-1) \qquad\qquad \frac{\partial \mathcal{N}_1}{\partial s} = \frac{1}{2}r(r-1)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_2}{\partial r} = \frac{1}{2}(2r+1)\frac{1}{2}s(s-1) \qquad\qquad \frac{\partial \mathcal{N}_2}{\partial s} = \frac{1}{2}r(r+1)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_3}{\partial r} = \frac{1}{2}(2r+1)\frac{1}{2}s(s+1) \qquad\qquad \frac{\partial \mathcal{N}_3}{\partial s} = \frac{1}{2}r(r+1)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_4}{\partial r} = \frac{1}{2}(2r-1)\frac{1}{2}s(s+1) \qquad\qquad \frac{\partial \mathcal{N}_4}{\partial s} = \frac{1}{2}r(r-1)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_5}{\partial r} = (-2r)\frac{1}{2}s(s-1) \qquad\qquad \frac{\partial \mathcal{N}_5}{\partial s} = (1-r^2)\frac{1}{2}(2s-1)$$

$$\frac{\partial \mathcal{N}_6}{\partial r} = \frac{1}{2}(2r+1)(1-s^2) \qquad\qquad \frac{\partial \mathcal{N}_6}{\partial s} = \frac{1}{2}r(r+1)(-2s)$$

$$\frac{\partial \mathcal{N}_7}{\partial r} = (-2r)\frac{1}{2}s(s+1) \qquad\qquad \frac{\partial \mathcal{N}_7}{\partial s} = (1-r^2)\frac{1}{2}(2s+1)$$

$$\frac{\partial \mathcal{N}_8}{\partial r} = \frac{1}{2}(2r-1)(1-s^2) \qquad\qquad \frac{\partial \mathcal{N}_8}{\partial s} = \frac{1}{2}r(r-1)(-2s)$$

$$\frac{\partial \mathcal{N}_9}{\partial r} = (-2r)(1-s^2) \qquad\qquad \frac{\partial \mathcal{N}_9}{\partial s} = (1-r^2)(-2s)$$

We go straight for the special case:

- $\rho_1 = \rho_2 = \rho_5$,

- $\rho_6 = \rho_8 = \rho_9$,

- $\rho_3 = \rho_4 = \rho_7$,

- $\theta_1 = \theta_4 = \theta_8$,

- $\theta_5 = \theta_7 = \theta_9$,

- $\theta_2 = \theta_3 = \theta_6$,

- $\frac{1}{2}(\theta_2 - \theta_1) = \frac{1}{2}(\theta_3 - \theta_4) = \frac{1}{2}(\theta_6 - \theta_8) = \tilde{\theta}$,

- $\frac{1}{2}(\rho_4 - \rho_1) = \frac{1}{2}(\rho_3 - \rho_2) = \frac{1}{2}(\rho_7 - \rho_5) = \bar{\rho}$

We start with

$$
\begin{aligned}
\frac{\partial \theta}{\partial r}(r,s) &= \sum_i \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\theta_i \\
&= \sum_{i=1,4,8} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\theta_i + \sum_{i=5,7,9} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\theta_i + \sum_{i=2,3,6} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)\theta_i \\
&= \theta_1 \sum_{i=1,4,8} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) + \theta_5 \sum_{i=5,7,9} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) + \theta_2 \sum_{i=2,3,6} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) \qquad (11.43)
\end{aligned}
$$

We find that

$$\sum_{i=1,4,8} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) = \frac{1}{2}(2r-1)\frac{1}{2}s(s-1) + \frac{1}{2}(2r-1)\frac{1}{2}s(s+1) + \frac{1}{2}(2r-1)(1-s^2)$$

$$= \frac{1}{2}(2r-1)\left[\frac{1}{2}s(s-1) + \frac{1}{2}s(s+1) + (1-s^2)\right]$$

$$= \frac{1}{2}(2r-1)\left[s^2 + (1-s^2)\right]$$

$$= \frac{1}{2}(2r-1)$$

$$\sum_{i=5,7,9} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) = -2r\frac{1}{2}s(s-1) - 2r\frac{1}{2}s(s+1) - 2r(1-s^2)$$

$$= -2r$$

$$\sum_{i=2,3,6} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) = \frac{1}{2}(2r+1)\frac{1}{2}s(s-1) + \frac{1}{2}(2r+1)\frac{1}{2}s(s+1) + \frac{1}{2}(2r+1)(1-s^2)$$

$$= \frac{1}{2}(2r+1)$$

Then

$$\frac{\partial \theta}{\partial r}(r,s) = \theta_1 \sum_{i=1,4,8} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) + \theta_5 \sum_{i=5,7,9} \frac{\partial \mathcal{N}_i}{\partial r}(r,s) + \theta_2 \sum_{i=2,3,6} \frac{\partial \mathcal{N}_i}{\partial r}(r,s)$$

$$= \frac{1}{2}(2r-1)\theta_1 - 2r\theta_5 + \frac{1}{2}(2r+1)\theta_2 \tag{11.44}$$

Using $\theta_5 = \frac{1}{2}(\theta_1 + \theta_2)$, then

$$\frac{\partial \theta}{\partial r}(r,s) = \frac{1}{2}(2r-1)\theta_1 - 2r\theta_5 + \frac{1}{2}(2r+1)\theta_2$$

$$= \frac{1}{2}(2r-1)\theta_1 - 2r\frac{1}{2}(\theta_1 + \theta_2) + \frac{1}{2}(2r+1)\theta_2$$

$$= \frac{1}{2}\left[(2r-1)\theta_1 - 2r\theta_1 - 2r\theta_2 + (2r+1)\theta_2\right]$$

$$= \frac{1}{2}(\theta_2 - \theta_1)$$

$$= \tilde{\theta} \tag{11.45}$$

Then we compute

$$\frac{\partial \theta}{\partial s}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\theta_i$$

$$= \sum_{i=1,4,8}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\theta_i + \sum_{i=5,7,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\theta_i + \sum_{i=2,3,6}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\theta_i$$

$$= \theta_1 \sum_{i=1,4,8}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \theta_5 \sum_{i=5,7,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \theta_2 \sum_{i=2,3,6}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) \tag{11.46}$$

570

We find that

$$\sum_{i=1,4,8} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = \frac{1}{2}r(r-1)\frac{1}{2}(2s-1) + \frac{1}{2}r(r-1)\frac{1}{2}(2s+1) + \frac{1}{2}r(r-1)(-2s)$$

$$= \frac{1}{2}r(r-1)\left[\frac{1}{2}(2s-1) + \frac{1}{2}(2s+1) - 2s\right]$$

$$= 0$$

$$\sum_{i=5,7,9} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = (1-r^2)\frac{1}{2}(2s-1) + (1-r^2)\frac{1}{2}(2s+1) + (1-r^2) + (-2s)$$

$$= 0$$

$$\sum_{i=2,3,6} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = \frac{1}{2}r(r+1)\frac{1}{2}(2s-1) + \frac{1}{2}r(r+1)\frac{1}{2}(2s+1)\frac{1}{2}r(r+1)(-2s)$$

$$= 0 \tag{11.47}$$

so that

$$\frac{\partial \theta}{\partial s}(r,s) = 0$$

At this stage we are reasonably sure that the components of the Jacobian $\boldsymbol{J}_{PL}$ we are computing for this Q_2 element will turn out to be the same at thos obtained for the Q_1 element. Just to be on the safe side, we soldier on:

$$\frac{\partial \rho}{\partial s}(r,s) = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\rho_i$$

$$= \sum_{i=1,2,5}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\rho_i + \sum_{i=6,8,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\rho_i + \sum_{i=3,4,7}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s)\,\rho_i$$

$$= \rho_1 \sum_{i=1,2,5}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \rho_6 \sum_{i=6,8,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \rho_3 \sum_{i=3,4,7}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) \tag{11.48}$$

with

$$\sum_{i=1,2,5}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = \frac{1}{2}r(r-1)\frac{1}{2}(2s-1) + \frac{1}{2}r(r+1)\frac{1}{2}(2s-1) + (1-r^2)\frac{1}{2}(2s-1)$$

$$= \frac{1}{2}(2s-1)\left[\frac{1}{2}r(r-1) + \frac{1}{2}r(r+1) + (1-r^2)\right]$$

$$= \frac{1}{2}(2s-1)$$

$$\sum_{i=6,8,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = \frac{1}{2}r(r+1)(-2s) + \frac{1}{2}r(r-1)(-2s) + (1-r^2)(-2s)$$

$$= -2s\left[\frac{1}{2}r(r+1) + \frac{1}{2}r(r-1) + (1-r^2)\right]$$

$$= -2s$$

$$\sum_{i=3,4,7}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) = \frac{1}{2}(2s+1)\left[\frac{1}{2}r(r+1) + \frac{1}{2}r(r-1) + (1-r^2)\right]$$

$$= \frac{1}{2}(2s+1)$$

In the end

$$
\begin{aligned}
\frac{\partial \rho}{\partial s}(r,s) &= \rho_1 \sum_{i=1,2,5}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \rho_6 \sum_{i=6,8,9}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) + \rho_3 \sum_{i=3,4,7}^{m} \frac{\partial \mathcal{N}_i}{\partial s}(r,s) \\
&= \rho_1 \frac{1}{2}(2s-1) + \rho_6(-2s) + \rho_3 \frac{1}{2}(2s+1)
\end{aligned}
$$

With $\rho_6 = \frac{1}{2}(\rho_1 + \rho_3)$ we find

$$
\begin{aligned}
\frac{\partial \rho}{\partial s}(r,s) &= \rho_1 \frac{1}{2}(2s-1) + \frac{1}{2}(\rho_1 + \rho_3)(-2s) + \rho_3 \frac{1}{2}(2s+1) \\
&= \frac{1}{2}\left[\rho_1(2s-1) + (\rho_1 + \rho_3)(-2s) + \rho_3(2s+1)\right] \\
&= \frac{1}{2}(\rho_3 - \rho_1) \\
&= \bar{\rho}
\end{aligned}
$$

Case closed, we find here too that

$$
\boldsymbol{J}_{PL} = \begin{pmatrix} \frac{\partial \theta}{\partial r} & \frac{\partial \rho}{\partial r} \\ \frac{\partial \theta}{\partial s} & \frac{\partial \rho}{\partial s} \end{pmatrix} = \begin{pmatrix} \tilde{\theta} & 0 \\ 0 & \bar{\rho} \end{pmatrix}
$$

Chapter 12

Solving the wave equation with the FEM

12.1 The wave equation

Disclaimer: what follows in this section is greatly inspired by `https://en.wikipedia.org/wiki/Wave_equation`

The wave equation is a second-order linear partial differential equation for the description of waves or standing wave fields such as mechanical waves (e.g. water waves, sound waves and seismic waves) or electromagnetic waves (including light waves). It arises in fields like acoustics, electromagnetism, and fluid dynamics.

The wave equation is a hyperbolic partial differential equation, and we will here focus on the scalar wave equation that represent the wave by a scalar $u(x, y, z, t)$. The derivation of the wave equation is presented there[1] and in virtually every physics textbook. The scalar wave equation is given by

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{12.1}$$

where c is a fixed non-negative real coefficient (the wave speed), u is a scalar field representing a displacement from rest situation, e.g. it could be gas pressure above or below normal, or the height of water in a pond above or below rest.

Note that sometimes you are likely to encounter different notations for the 1st and 2nd-order derivatives:

$$u_t = \partial_t u = \frac{\partial u}{\partial t} \qquad u_{xx} = \frac{\partial^2 u}{\partial x^2}$$

so that Eq. (12.1) then becomes

$$u_{tt} = c^2 (u_{xx} + u_{yy} + u_{zz})$$

The equation states that at any given instance, at any given point, the acceleration of the displacement is proportional to the way the displacement's changes are squashed up in the surrounding area. In other words, a more pointy displacement gets pushed back more forcefully.

This equation can also be written

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \Delta u \tag{12.2}$$

or even

$$\Box u = 0 \quad \text{with} \quad \Box = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} - \Delta$$

---

[1] `https://en.wikipedia.org/wiki/Wave_equation`

in its most compact form, where $\square$ is the d'Alembert operator[2].

Solving the wave equation can be quite complicated. Still, it can be analyzed as a linear combination of simple solutions that are sinusoidal plane waves with various directions of propagation and wavelengths but all with the same propagation speed c. This analysis is possible because the wave equation is linear and homogeneous, so that any multiple of a solution is also a solution, and the sum of any two solutions is again a solution. This property is called the superposition principle in physics.

The wave equation alone does not specify a physical solution; a *unique* solution is usually obtained by setting a problem with further conditions, such as initial conditions, which prescribe the amplitude and phase of the wave. Another important class of problems occurs in enclosed spaces specified by boundary conditions, for which the solutions represent standing waves, or harmonics, analogous to the harmonics of musical instruments.

12.2 Solving the one-dimensional wave equation

The one-dimensional wave equation simply writes

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad \text{or} \qquad u_{tt} = c^2 u_{xx}$$

with $c > 0$, to be solved on the domain $x \in [0, L_x]$. The function u can be the displacement of a string or membrane of a drum, the height of a water wave or air pressure for sound waves, dependent on the field the equation is used in.

The constant c in the above equations defines the speed at which the wave "moves". To see this we can look at the one dimensional wave equation for which, if we have no other conditions, $u(x,t) = f(x + ct)$ and $u(x,t) = f(x - ct)$ are solutions if the function f is twice differentiable. These two solutions have a constant shape which move in time with speed c to the left and right respectively. A general solution to the one- dimensional wave equation was actually derived by Jean le Rond d'Alembert [3] (1717-1783) and is given by $u(x,t) = f(x - ct) + g(x + ct)$ being the sum of a left and right moving wave.

12.2.1 Initial and boundary conditions

Simply put, we need to specify the state of the system u at the beginning, and since the domain in which the wave propagates is likely to be finite, also the boundary conditions.

For simplicity we will set

$$
\begin{aligned}
u(x = 0, t) &= 0 & \forall t \in \mathbb{R}^+ \\
u(x = L_x, t) &= 0 & \forall t \in \mathbb{R}^+ \\
u(x, 0) &= u_0(x) & \forall x \in [0, L_x] \\
u_t(x, 0) &= u_1(x) & \forall x \in [0, L_x]
\end{aligned}
\tag{12.3}
$$

Here, $u(x,t)$ represents the the displacement of a string (in one-dimension) or membrane of a drum (in two-dimensions) that is held fixed at the boundary.

12.2.2 A simple example: a standing wave

Let us set $L_x = 1$ and $c = 1$. Then one of the solutions is given by

$$x(x,t) = \cos(2\pi t)\sin(2\pi x)$$

---

[2] https://en.wikipedia.org/wiki/D'Alembert_operator
[3] https://en.wikipedia.org/wiki/Jean_le_Rond_d'Alembert

which is periodic in time with period 1 and is called a standing wave. We also have

$$u_t(x,t) = -2\pi \sin(2\pi t)\sin(2\pi x)$$

so that

$$
\begin{align}
u_0(x) &= \sin(2\pi x) \tag{12.4}\\
u_1(x) &= 0 \tag{12.5}
\end{align}
$$

and the boundary conditions are respected: $u(0,t) = u(1,t) = 0 \quad \forall t$.
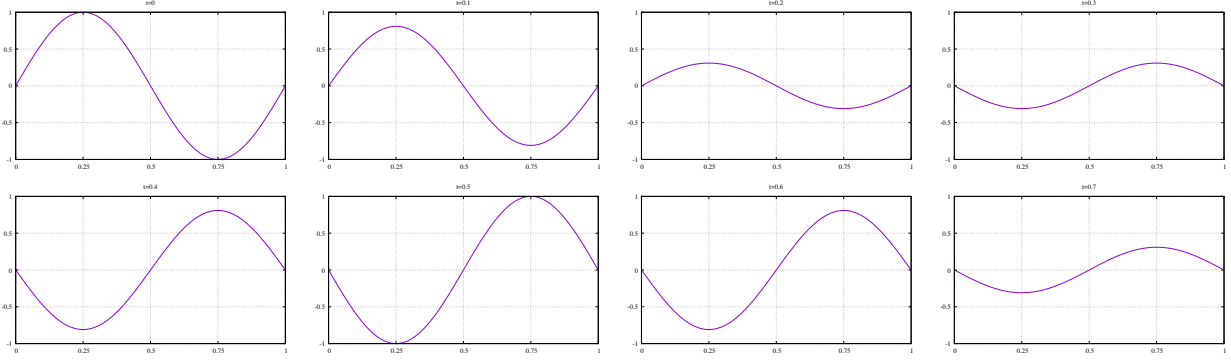
Solution at different times.

12.2.3 A few words about the FEM for the (1D) wave equation

The finite element method is a numerical method generally used to solve differential equations with boundary conditions. The concept of the finite element method is to divide the domain into finitely many smaller areas called elements and approximate the solution of the differential equation on these elements using a suitable set of basis functions. The result is a system of equations which can either be solved directly or by another numerical method.

The finite element method has a few advantages over other numerical methods for solving differential equations (like the finite difference method). Firstly, the finite element method approximates the solution for almost the whole domain, while most other methods only approximate the solution at a discrete set of points. Secondly, with the finite element method we have more freedom over the discretization of the do- main. We have the choice to increase precision at parts of the domain by using smaller elements and to decrease the precision at parts by using larger elements. Thirdly we can choose different basis functions on the elements to increase or decrease the precision at different parts of the domain or choose basis functions which better approximate the actual solution, for example by using trigonometric functions instead of polynomials.

12.2.4 Weak formulation

First we will derive a different but equivalent formulation for the one-dimensional wave equation. Let $v(x)$ be a differentiable functions such that $v(0) = v(1) = 0$ (i.e. zero on the boundaries of the domain). By multiplying the wave equation with this function and integrating it over the domain we get following expression:

$$\int_0^L v(x)\left(\frac{\partial^2 u}{\partial t^2}(x,t) - c^2\frac{\partial^2 u}{\partial x^2}(x,t)\right)dx = 0$$

Let us recall the integration by parts formulation:

$$\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx$$

and use it on the equation above (only for the space derivative term):

$$\int_0^L v(x)\frac{\partial^2 u}{\partial t^2}(x,t)dx - c^2 \int_0^L v(x)\left(\frac{\partial^2 u}{\partial x^2}(x,t)\right)dx = 0$$

$$\int_0^L v(x)\frac{\partial^2 u}{\partial t^2}(x,t)dx - c^2 \left[v(x)\frac{\partial u}{\partial x}(x,t)\right]_0^L + c^2 \int_0^L \frac{\partial v}{\partial x}(x,t)\frac{\partial u}{\partial x}(x,t)dx = 0$$

Since v is zero on the boundary the middle term is then zero and we obtain:

$$\int_0^L v(x)\frac{\partial^2 u}{\partial t^2}(x,t)dx + c^2 \int_0^L \frac{\partial v}{\partial x}(x,t)\frac{\partial u}{\partial x}(x,t)dx = 0 \tag{12.6}$$

or,

$$a(u(x,t),v(x)) = 0$$

with

$$a(u(x,t),v(x)) := \int_0^L v(x)\frac{\partial^2 u}{\partial t^2}(x,t)dx + c^2 \int_0^L \frac{\partial v}{\partial x}(x)\frac{\partial u}{\partial x}(x,t)dx \tag{12.7}$$

For the integrals above to exist we require certain properties of u, v and their derivatives. For this we will use the L_2-norm on domain $D = [0, L]$. For a function f this is defined by

$$\|f\|_2 := \int_D |f(x)|^2 dx$$

Using this norm a function f is called square-integrable if the L_2-norm of f is finite or more specifically:

$$\|f\|_2 < \infty$$

The set of all square-integrable functions is denoted by L_2. An important property of square-integrable functions is that the integral of the product of two square-integrable functions over the domain D is also finite. This property is important for the existence of the integrals above.

We want the functions u, v and their first derivatives with respect to x to be square-integrable. The set of functions which are square-integrable and whose first derivative is square-integrable is called a first order Sobolev space and can be denoted by H^1. The set of functions which are also zero at the boundary of the domain is denoted by H_0^1.

We now want to find a function u which is in H_0^1 for each moment in time t such that for every function v in H_0^1 the equations of Eq. (12.6) holds. This gives rise to the weak formulation of our one-dimensional wave equation, which can now be stated as follows

Definition 1.1.1 (Weak formulation). Find $u(x,t)$ such that for all $t \in \mathbb{R}$ it holds that $u(\cdot, t) \in H_0^1$ and for all $v \in H_0^1$

$$a(u(x,t),v(x)) = 0 \tag{12.8}$$

and the initial boundary conditions (12.3) holds.

This is called the weak formulation since the function u only has to verify Eq. (12.8) with respect to certain functions v called test functions. Except for the derivation of the weak formulation above the rigorous connection between the weak formulation and the "strong formulation" (i.e. the wave equation) is beyond the scope of this work.

12.2.5 Nodes and Elements

We will now divide the interval $[0, L_x]$ into $nelx$ elements of equal size separated by $nnx = nelx + 1$ nodes. The size of each element is given by $h = L_x/nelx$, so that the node coordinates can be defined as

$$x_i := (i - 1)h \qquad \text{for} \quad i = 1, ...nnx$$

The nodes and elements are illustrated in the following figure:
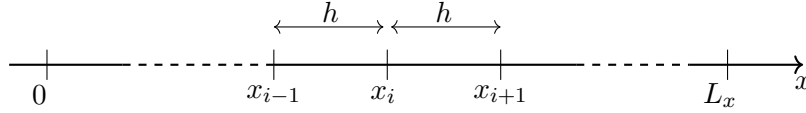
12.2.6 Basis functions

Instead of searching for a function u in H_0^1 such that for all test functions v in H_0^1 the equation of the weak formulation holds, we are going to search in a finite-dimensional subspace of H^1. This space will consist of those functions in H^1 which are piecewise polynomial on each element. We will restrict ourselves to piecewise linear functions for simplicity and we will denote this finite-dimensional set of functions as S^1.

The set S^1 will be defined as the functions v that can be written as a linear combination of finitely many basis functions and that are zero at the boundaries. We will take nnx basis functions, one for each node, and call these $\mathcal{N}_i(x)$ for $i = 1, ..., nnx$. Then the functions v in S^1 can be written as:

$$v(x) = \sum_{i=1}^{nnx} \mathcal{N}_i(x)\, v_i \qquad \text{with} \quad v_i \in \mathbb{R}, \quad v(0) = v(L_x) = 0. \tag{12.9}$$

The basis functions can be written (for $i > 1$ and $i < nnx$):

$$
\begin{aligned}
\mathcal{N}_i(x) &= 0 && \text{if } x < x_{i-1} \\
\mathcal{N}_i(x) &= \frac{x - x_{i-1}}{h} && \text{if } x \in [x_{i-1}, x_i] \\
\mathcal{N}_i(x) &= \frac{x_{i+1} - x}{h} && \text{if } x \in [x_i, x_{i+1}] \\
\mathcal{N}_i(x) &= 0 && \text{if } x > x_{i+1}
\end{aligned}
$$

as shown in the following figure:
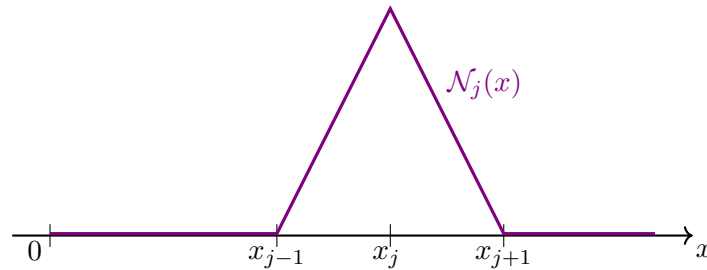
Basis functions for $j = 1$ and $j = nnx$ only contain the relevant half of the hat functions defined above. Let us recall that an important property of these basis functions is that $\mathcal{N}_i(x)$ is precisely one at node x_i and zero on every other node.

12.2.7 Finite element solution

Using the just defined basis functions we can define a finite element solution and finite element test functions as follows

$$U(x,t) := \sum_{i=1}^{nnx} \mathcal{N}_i(x)U_i(t)$$

$$V(x) := \sum_{i=1}^{nnx} \mathcal{N}_i(x)V_i$$

Definition 1.4.1 (Finite element formulation). Find $U(x,t)$ such that for all $t \in \mathbb{R}$ it holds that $U(\cdot,t) \in S^1$ and for all $V \in S^1$:

$$a(U(x,t),V(x)) = 0$$

and the following initial boundary conditions hold

$$
\begin{array}{ll}
U(x=0,t) = 0 & \forall t \in \mathbb{R}^+ \\
U(x=L_x,t) = 0 & \forall t \in \mathbb{R}^+ \\
U(x,0) = U_0(x) & \forall x \in [0, L_x] \\
U_t(x,0) = U_1(x) & \forall x \in [0, L_x]
\end{array}
\tag{12.10}
$$

Note that because of the form U takes, it does not have to hold that $U_0(x)$ equals $u_0(x)$.

For the finite element solution it is enough to test it only against the basis functions. This follows from the linearity of $a(u,v)$ in the second variable and therefore it holds that

$$a(U(x,t),V(x)) = a\left(U(x,t), \sum_{i=1}^{nnx} \mathcal{N}_j(x)V_j\right) = \sum_{i=j}^{nnx} V_j \, a\left(U(x,t), \mathcal{N}_j(x)\right)$$

Thus, if it holds for all of the basis functions $\mathcal{N}_j$, it also holds for all functions $V \in S^1$.

For each basis function $\mathcal{N}_j(x)$ that is not on the edge, i.e. $j = 2, ..., nnx - 1$ we can write using the definition of the bilinear form $a(\cdot,\cdot)$ (see Eq. (12.7)):

$$
\begin{aligned}
0 &= a\left(U(x,t), \mathcal{N}_j(x)\right) \\
&= \int_0^L \mathcal{N}_j(x)\frac{\partial^2 U}{\partial t^2}(x,t)dx + c^2 \int_0^L \frac{\partial \mathcal{N}_j}{\partial x}(x)\frac{\partial U}{\partial x}(x,t)dx \\
&= \int_0^L \mathcal{N}_j(x)\frac{\partial^2 \left(\sum_{i=1}^{nnx}\mathcal{N}_i(x)U_i(t)\right)}{\partial t^2}(x,t)dx + c^2 \int_0^L \frac{\partial \mathcal{N}_j}{\partial x}(x)\frac{\partial \left(\sum_{i=1}^{nnx}\mathcal{N}_i(x)U_i(t)\right)}{\partial x}(x,t)dx \\
&= \sum_{i=1}^{nnx} \frac{d^2 U_i(t)}{dt^2} \int_0^L \mathcal{N}_j(x)\mathcal{N}_i(x)dx + c^2 \sum_{i=1}^{nnx} U_i(t) \int_0^L \frac{d\mathcal{N}_j}{dx}(x)\frac{d\mathcal{N}_i}{dx}(x)dx \\
&= \sum_{i=1}^{nnx} \ddot{U}_i(t) \underbrace{\int_0^L \mathcal{N}_j(x)\mathcal{N}_i(x)dx}_{M_{ij}} + c^2 \sum_{i=1}^{nnx} U_i(t) \underbrace{\int_0^L \frac{d\mathcal{N}_j}{dx}(x)\frac{d\mathcal{N}_i}{dx}(x)dx}_{K_{ij}}
\end{aligned}
$$

$$\tag{12.11}$$

This results in $nnx - 2$ linear ordinary differential equations. For the two basis functions left out we can use the boundary conditions imposed on U. By defining the vector $\vec{\mathcal{U}} = (U_1(t), U_2(t), ...U_{nnx}(t))$ we can use these nnx linear differential equations to write the following ordinary differential equation

$$\mathbb{M} \cdot \ddot{\vec{\mathcal{U}}}(t) + c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t) = \vec{0} \qquad (12.12)$$

The values of M_{ij} and K_{ij} can easily be calculated by splitting their integrals over all the elements like

$$\int_0^{L_x} f(x)dx = \sum_{e=1}^{nelx} \int_{x_e}^{x_{e+1}} f(x)dx$$

As the basis functions only have a maximum of two elements on which they are non-zero, the values of Ti,j and Si,j can be calculated by just integrating over the few elements on which both basis functions are non-zero. This allows us to calculate the following coefficients

$$
\begin{aligned}
M_{i,i} &= \int_0^L \mathcal{N}_i(x)\mathcal{N}_i(x)dx = \int_{x_{i-1}}^{x_{i+1}} \mathcal{N}_i(x)\mathcal{N}_i(x)dx = \frac{2}{3}h \\
M_{i,i-1} &= \int_0^L \mathcal{N}_i(x)\mathcal{N}_{i-1}(x)dx = \int_{x_{i-1}}^{x_i} \mathcal{N}_i(x)\mathcal{N}_{i-1}(x)dx = \frac{1}{6}h \\
K_{i,i} &= \int_0^L \mathcal{N}_i'(x)\mathcal{N}_i'(x)dx = \int_{x_{i-1}}^{x_{i+1}} \mathcal{N}_i'(x)\mathcal{N}_i'(x)dx = \frac{2}{h} \\
K_{i,i-1} &= \int_0^L \mathcal{N}_i'(x)\mathcal{N}_{i-1}'(x)dx = \int_{x_{i-1}}^{x_i} \mathcal{N}_i'(x)\mathcal{N}_{i-1}'(x)dx = -\frac{1}{h} \qquad (12.13)
\end{aligned}
$$

These integrals can easily be computed by hand in the 1D case but we will see that in most 2D,3D cases this is not the case and we must resort to numerical quadrature.

12.2.8 Time discretisation

We can use a second-order approximation for the second-order derivative:

$$\ddot{\vec{\mathcal{U}}}(t) \simeq \frac{\vec{\mathcal{U}}(t+\delta t) - 2\vec{\mathcal{U}}(t) + \vec{\mathcal{U}}(t-\delta t)}{\delta t^2}$$

which we can substitute in Eq. (12.12):

$$\mathbb{M} \cdot \frac{\vec{\mathcal{U}}(t+\delta t) - 2\vec{\mathcal{U}}(t) + \vec{\mathcal{U}}(t-\delta t)}{\delta t^2} + c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t) = \vec{0}$$

Note that in the equation above the vector $\vec{\mathcal{U}}$ multiplying $\mathbb{K}$ is evaluated at time t. We will come back to this very shortly.

$$\mathbb{M} \cdot (\vec{\mathcal{U}}(t+\delta t) - 2\vec{\mathcal{U}}(t) + \vec{\mathcal{U}}(t-\delta t)) = -c^2\delta t^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t)$$

leading to finally write (method # 1):

$$\mathbb{M} \cdot \vec{\mathcal{U}}(t+\delta t) = \underbrace{[2\mathbb{M} - c^2\delta t^2\mathbb{K}] \cdot \vec{\mathcal{U}}(t) - \mathbb{M} \cdot \vec{\mathcal{U}}(t-\delta t)}_{\vec{b}}$$

which is of the form

$$\mathcal{A} \cdot \vec{\mathcal{U}}(t+\delta t) = \vec{b}$$

Alternatively we would have used an implicit approach by writing

$$\mathbb{M} \cdot \frac{\vec{\mathcal{U}}(t+\delta t) - 2\vec{\mathcal{U}}(t) + \vec{\mathcal{U}}(t-\delta t)}{\delta t^2} + c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t+\delta t) = \vec{0},$$

leading to write:

$$(\mathbb{M} + c^2 \delta t^2 \mathbb{K}) \cdot \vec{\mathcal{U}}(t + \delta t) = \underbrace{\mathbb{M} \cdot (2\vec{\mathcal{U}}(t) - \vec{\mathcal{U}}(t - \delta t))}_{\vec{b}}$$

Given the presence of the terms $\vec{\mathcal{U}}(t)$ and $\vec{\mathcal{U}}(t - \delta t)$ in the equations above, we can only solve the equation for $t \geq 2\delta t$ (assuming the time step δt remains constant):

$$\mathbb{M} \cdot \vec{\mathcal{U}}(2\delta t) = [2\mathbb{M} - c^2 \delta t^2 \mathbb{K}] \cdot \vec{\mathcal{U}}(\delta t) - \mathbb{M} \cdot \vec{\mathcal{U}}(0)$$

The value of $\vec{\mathcal{U}}(0)$ is known as it is the field at $t = 0$ (initial condition). However, quid of $\vec{\mathcal{U}}(\delta t)$? One way to avoid this problem is by using a 1st order Taylor expansion of this term:

$$\vec{\mathcal{U}}(\delta t) = \vec{\mathcal{U}}(0) + \delta t \, \dot{\vec{\mathcal{U}}}(0)$$

and since we know both terms in the rhs then all is well.

Another approach can be taken (method # 2): let us define

$$\vec{\mathcal{V}} = \frac{\vec{\mathcal{U}}(t + \delta t) - 2\vec{\mathcal{U}}(t) + \vec{\mathcal{U}}(t - \delta t)}{\delta t^2}$$

Then one needs to solve two equations one after the other:

$$\begin{aligned}
\mathbb{M} \cdot \vec{\mathcal{V}} &= -c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t) \\
\vec{\mathcal{U}}(t + \delta t) &= \delta t^2 \, \vec{\mathcal{V}} + 2\vec{\mathcal{U}}(t) - \vec{\mathcal{U}}(t - \delta t)
\end{aligned} \tag{12.14}$$

There is yet another option (method # 3). Let us define this time $\vec{\mathcal{W}}(t) = \dot{\vec{\mathcal{U}}}(t)$. Then Eq. (12.12) can be written

$$\dot{\vec{\mathcal{W}}}(t) = \ddot{\vec{\mathcal{U}}} = -c^2 \mathbb{M}^{-1} \cdot \mathbb{K} \cdot \vec{\mathcal{U}}(t)$$

Then we must solve the coupled set of equations:

$$\dot{\vec{\mathcal{U}}}(t) = \vec{\mathcal{W}}(t) \tag{12.15}$$

$$\dot{\vec{\mathcal{W}}}(t) = -c^2 \mathbb{M}^{-1} \cdot \mathbb{K} \cdot \vec{\mathcal{U}}(t) \tag{12.16}$$

with initial values

$$\vec{\mathcal{U}}(0) = \vec{\mathcal{U}}_0 \tag{12.17}$$

$$\vec{\mathcal{W}}(0) = \dot{\vec{\mathcal{U}}}_0 \tag{12.18}$$

We can also write (to first order)

$$\frac{\vec{\mathcal{U}}(t + dt) - \vec{\mathcal{U}}(t)}{dt} = \vec{\mathcal{W}}(t) \tag{12.19}$$

$$\mathbb{M} \cdot \frac{\vec{\mathcal{W}}(t + dt) - \vec{\mathcal{W}}(t)}{dt} = -c^2 \, \mathbb{K} \cdot \vec{\mathcal{U}}(t) \tag{12.20}$$

and finally

$$\vec{\mathcal{U}}(t + dt) = \vec{\mathcal{U}}(t) + dt\,\dot{\vec{\mathcal{U}}}(t) \tag{12.21}$$

$$\mathbb{M} \cdot \frac{\dot{\vec{\mathcal{U}}}(t + dt) - \dot{\vec{\mathcal{U}}}(t)}{dt} = -c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t) \tag{12.22}$$

The first equation is straightforward and we then solve the second equation in a similar was as in method 2 to obtain $\dot{\vec{\mathcal{U}}}(t + dt)$.

Concretely, it goes as follows. Let us define $\vec{\mathcal{R}} = (\dot{\vec{\mathcal{U}}}(t + dt) - \dot{\vec{\mathcal{U}}}(t))/dt$. Then the second equation is $\mathbb{M} \cdot \vec{\mathcal{R}} = -c^2\,\mathbb{K} \cdot \vec{\mathcal{U}}(t)$ and the algorithm goes as follows:

1. $\vec{\mathcal{U}}(t + dt) = \vec{\mathcal{U}}(t) + dt\,\dot{\vec{\mathcal{U}}}(t)$

2. solve $\mathbb{M} \cdot \vec{\mathcal{R}} = -c^2\,\mathbb{K} \cdot \vec{\mathcal{U}}(t)$

3. recover $\dot{\vec{\mathcal{U}}}(t + dt) = \dot{\vec{\mathcal{U}}}(t) + dt\,\vec{\mathcal{R}}$

This method is different than the other two because it keeps $\vec{\mathcal{U}}$ and its time-derivative as unknowns.

Note that in all three cases the FE matrix is simply the mass matrix, which is good news since it tends to be nicely conditioned and symmetric.

This is all implemented in STONE 164.

12.2.9 The case of non-equidistant nodes and/or higher order elements

In this section we will generalize the finite element method by first allowing irregular elements, i.e. elements of different size:
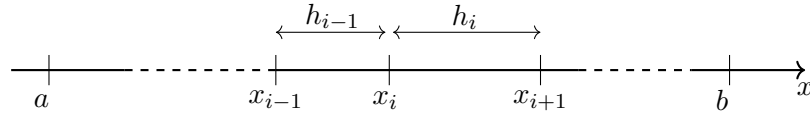
We divide an interval [a,b] with $a < b$ into $nelx$ elements separated by nodes. The nodes x_i with $i = 1, ...nnx$ are chosen such that $a = x_1 < x_2 < ... < x_{nnx-1} < x_{nnx} = b$ and the elements can then defined as

$$e_i = [x_i, x_{i+1}] \qquad \text{for} \quad i = 1, ..., nnx$$

with its size $h_i := x_{i+1} - x_i$.

We have now two choices: either we define a so-called reference element (sometimes also called canonical element) and establish the mapping from any real element to the reference one, later to be used in the formulation of the integrals, or we express the basis functions locally for each element and proceed by using these in the integrals.

For example the basis functions would then write

$$\begin{aligned}
\mathcal{N}_i(x) &= 0 & &\text{if } x < x_{i-1} \\
\mathcal{N}_i(x) &= \frac{x - x_{i-1}}{x_i - x_{i-1}} & &\text{if } x \in [x_{i-1}, x_i] \\
\mathcal{N}_i(x) &= \frac{x_{i+1} - x}{x_{i+1} - x_i} & &\text{if } x \in [x_i, x_{i+1}] \\
\mathcal{N}_i(x) &= 0 & &\text{if } x > x_{i+1}
\end{aligned}$$

as shown in the following figure:
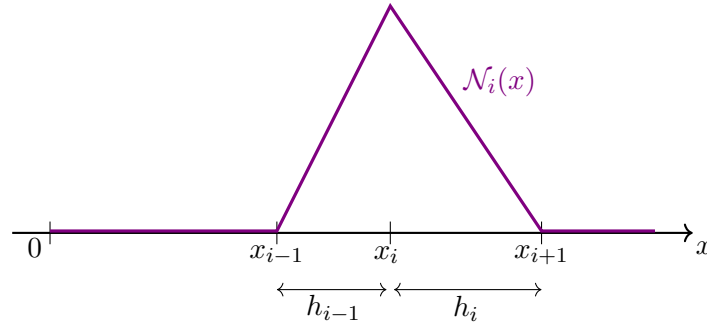
Otherwise we would normally define the reference element $r = [-1, +1]$ and the basis functions inside the element attached to each node would be given by

$$\mathcal{N}_1(r) = \frac{1}{2}(1 - r) \tag{12.23}$$

$$\mathcal{N}_2(r) = \frac{1}{2}(1 + r) \tag{12.24}$$

We have of course already encountered these in Section **??**.

The topic of higher-order elements has been covered in Section **??** and poses no real difficulty. For example, 2nd-order elements would require an additional node in the middle of each element while the basis functions are defined on the reference element in section REF ...

12.2.10   Conservation of energy

The concept of conservation of energy is used all over physics and states most of the time that the total amount of energy in a closed systems is constant. This conservation can also hold for the one-dimensional wave equation and can be formulated as

Theorem   For a solution u of the one-dimensional wave equation there exists a constant $C > 0$, independent of time t, such that

$$\left\| \frac{\partial u}{\partial t}(\cdot, t) \right\|^2 + c^2 \left\| \frac{\partial u}{\partial x}(\cdot, t) \right\|^2 = C \tag{12.25}$$

where $\| \cdot \|$ is the L_2-norm over the domain.

Proof: Let $v(x, t) = \frac{\partial u}{\partial t}(x, t)$ be the test function in the weak formulation of the one-dimensional wave equation. Using Eq. (12.7), this gives

$$
\begin{aligned}
0 &= \int_D v(x,t) \frac{\partial^2 u}{\partial t^2}(x,t)dx + c^2 \int_D \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}dx \\
&= \int_D \frac{\partial u}{\partial t}(x,t)\frac{\partial^2 u}{\partial t^2}(x,t)dx + c^2 \int_D \frac{\partial u}{\partial x}\frac{\partial}{\partial x}\frac{\partial u}{\partial t}(x,t)dx \\
&= \int_D \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 dx + c^2 \int_D \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right)^2 dx \\
&= \frac{1}{2}\frac{\partial}{\partial t}\int_D \left(\frac{\partial u}{\partial t}\right)^2 dx + c^2\frac{1}{2}\frac{\partial}{\partial t}\int_D \left(\frac{\partial u}{\partial x}\right)^2 dx \\
&= \frac{1}{2}\frac{\partial}{\partial t}\left(\left\|\frac{\partial u}{\partial t}(\cdot,t)\right\|^2 + c^2\left\|\frac{\partial u}{\partial x}(\cdot,t)\right\|^2\right)
\end{aligned} \tag{12.26}
$$

582

which implies Eq. (12.25). This implies conservation of energy for the finite element solution of the wave equation.

From the above formulas we can define an energy function as $E(u,t) = \|\dot{u}\|^2 + c^2\|\partial_x u\|^2$, such that for a solution u of the weak or finite element formulation there exists a constant $C > 0$ such that $E(u,t) = C$ for all t.

Considering now the discrete solution U, the first error we can notice is that the initial values of the wave equation $u(x,0)$ and $\partial_t u(x,0)$ and its discretized value for the finite element formulation $U(x,0)$ and $\partial_t U(x,0)$ do not necessarily have the same energy, or more specific, $E(u,0)$ and $E(U,0)$ do not have to be equal. But if we make the elements smaller and use higher order polynomials for approximation, we can better approximate the original initial value with U and $\partial_t U$ and therefore better approximate the energy.

To analyse what happens to the energy when we discretize in time the system of ordinary differential equations of Eq. (12.12) we can use a more efficient way of calculating the energy. We can get this by putting the finite element $U(x,t) = \sum_{i=1}^{nnx} \mathcal{N}_i(x) U_i(t)$ in the energy function, resulting in the following equations:

$$
\begin{aligned}
E(U,t) &= \left\|\frac{\partial u}{\partial t}(\cdot,t)\right\|^2 + c^2 \left\|\frac{\partial u}{\partial x}(\cdot,t)\right\|^2 \\
&= \int_D \left(\sum_{i=1}^{nnx} \mathcal{N}_i(x)\dot{U}_i(t)\right)^2 dx + c^2 \int_D \left(\sum_{i=1}^{nnx} \frac{d\mathcal{N}_i}{dx}(x)U_i(t)\right)^2 dx \\
&= \int_D \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} \mathcal{N}_i(x)\mathcal{N}_j(x)\dot{U}_i(t)\dot{U}_j(t)dx + c^2 \int_D \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} \frac{d\mathcal{N}_i}{dx}(x)\frac{d\mathcal{N}_j}{dx}(x)U_i(t)U_j(t)dx \\
&= \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} \dot{U}_i(t)\dot{U}_j(t)\int_D \mathcal{N}_i(x)\mathcal{N}_j(x)dx + c^2 \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} U_i(t)U_j(t)\int_D \frac{d\mathcal{N}_i}{dx}(x)\frac{d\mathcal{N}_j}{dx}(x)dx \\
&= \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} \dot{U}_i(t)\dot{U}_j(t)M_{ij} + c^2 \sum_{i=1}^{nnx}\sum_{j=1}^{nnx} U_i(t)U_j(t)K_{ij}dx \\
&= \vec{\mathcal{U}}^T \cdot \mathbb{M} \cdot \vec{\mathcal{U}} + c^2 \vec{\mathcal{U}}^T \cdot \mathbb{K} \cdot \vec{\mathcal{U}} \qquad\qquad (12.27)
\end{aligned}
$$

We see that methods 1&2 might not be convenient when it comes to energy calculations since the time derivative of the $\vec{\mathcal{U}}$ vector does not appear explicitly, while it does in method 3.

Because we know that real solutions of the wave equation and finite element formulation need to have a constant energy, we can use the energy function as a measure of error over time when we have all the values U_i and $\dot{U}_i$ at a moment in time, like for the third method. We can even use this method this analyse numerical solutions for which we don't have the analytical solutions.

12.2.11 time step value & CFL condition

:

12.3 Solving the two-dimensional wave equation

It simply writes

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

or also sometimes

$$u_{tt} = c^2 (u_{xx} + u_{yy})$$

Here to initial conditions and boundary conditions must be provided.

... include here weak form derivation and discretisation ...

We arrive again at the following equation

$$\mathbb{M} \cdot \vec{\ddot{\mathcal{U}}}(t) + c^2 \mathbb{K} \cdot \vec{\mathcal{U}}(t) = \vec{0} \tag{12.28}$$

At this stage a decision must be made, i.e. whether quadrilaterals or triangles are used, and whether they are linear, quadratic, etc ...

Chapter 13

Computing derivatives of FEM primary variables

Chapter 14

The Discontinuous Galerkin Finite Element Method (DG-FEM)

What is DG?

- it is a variant of the SG ("Standard Galerkin FEM")[1]

- SG-FEM requires continuity of the solution along element interfaces (edges).

- DG-FEM does not require continuity of the solution along edges.

- DG methods have more degrees of freedom than SG methods.

- DG-FEM shares some properties with FVM

Various books about DG-FEM

- *Discontinuous Galerkin Methods. Theory, Computation and Applications* by Cockburn, Karniadakis and Shu [285]

- *Mathematical Aspects of Discontinuous Galerkin Methods* by Di Pietro and Ern [1061]

- *Discontinuous Galerkin Methods. Analysis and Applications to Compressible Flow* by Dolejsi and Feistauer [359]

- *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations* by Riviére [1145]

- *Discontinuous finite elements in fluid dynamics and heat transfer* by Li [825]

- *Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications* by Hesthaven & Warbuton [601]

---

[1]Some authors use the acronym CG for Continuous Galerkin but since the Conjugate Gradient solver acronym CG is very much present in FE codes it can be confusing so we use here SG instead.

DG flavors There are many different flavours of the Discontinuous Galerkin Finite Element Method:

- **HDG**: Hybridizable DG [281, 286, 994, 995, 993]

- **IPG**: Interior Penalty G [945, 946]

- **IIPG**: Incomplete Interior Penalty G [360]

- **SIPG**: Symmetric Interior Penalty G [125, 1214]

- **LDG**: Local DG [232, 283, 227, 282]

Pro and cons for DG-FEM versus SG-FEM

- Assembly of stiffness matrix is easier to implement (ref?).

- Refinement of triangles is easier to implement (ref?).

- DG methods can easily handle adaptivity strategies since refinement or unrefinement of the grid can be achieved without taking into account the continuity restrictions typical of conforming finite element methods. Moreover, the degree of the approximating polynomial can be easily changed from one element to the other. [284]

- DG methods can support high order local approximations that can vary nonuniformly over the mesh.

- DG methods are readily parallelizable. Since the elements are discontinuous, the mass matrix is block diagonal and since the size of the blocks is equal to the number of degrees of freedom inside the corresponding elements, the blocks can be inverted by hand once and for all.[284]

The DG-FEM in geodynamics   This method has not been used extensively in geodynamics with (so-far) two noticeable exceptions:

- Lehmann *et al.* , Comparison of continuous and discontinuous Galerkin approaches for variable-viscosity Stokes flow (2015) [807]

- He *et al.* , A discontinuous Galerkin method with a bound preserving limiter for the advection of non-diffusive fields in solid Earth geodynamics (2017) [589]

- Puckett *et al.* , New numerical approaches for modeling thermochemical convection in a compositionally stratified fluid (2018) [1085]

14.1   First-order advection ODE in 1D

What follows is borrowed from the book *Discontinuous finite elements in fluid dynamics and heat transfer* by Ben Q. Li [825].

To illustrate the basic ideas of the discontinuous finite element method, we consider a simple, one-dimensional, first order differential equation with u specified at one of the boundaries:

$$\frac{du}{dx} + g = 0 \qquad x \in [a,b] \qquad \text{and} \qquad u(x=a) = u_a \tag{14.1}$$

where g is a constant (for simplicity). The domain is discretized such that : $\Omega_j = [x_j, x_{j+1}]$ with $j = 1, 2, ..., nel$. Then, integrating the above equation over the element j with respect to a weighting function $f(x)$

$$\int_{x_j}^{x_{j+1}} \left(\frac{du}{dx} + g \right) f(x) dx = 0 \tag{14.2}$$

Remembering that $\int_c^d u(x)v'(x)dx = [u(x)v(x)]_c^d - \int_c^d u'(x)v(x)dx$, we can now perform an integration by parts on the differential operator and we obtain:

$$[u(x)f(x)]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} \left(u\frac{df}{dx} - gf(x) \right) dx = 0 \tag{14.3}$$

or,

$$u(x_{j+1})f(x_{j+1}) - u(x_j)f(x_j) - \int_{x_j}^{x_{j+1}} \left(u\frac{df}{dx} - gf(x) \right) dx = 0 \tag{14.4}$$

On the element Ω_j the function u is approximated by $u_h \in H$, H being an appropriate function space of finite dimension, and f by f_h taken from the same function space as u_h. Upon substituting (u_h, f_h) for (u, f) in the equation above, we have the discontinuous Galerkin finite element formulation:

$$u_h(x_{j+1})f_h(x_{j+1}) - u_h(x_j)f_h(x_j) - \int_{x_j}^{x_{j+1}} \left(u_h\frac{df_h}{dx} - gf_h(x) \right) dx = 0 \tag{14.5}$$

In the continuous finite element approach, the field variable u_h is forced to be continuous across the boundary. The essential idea for the discontinuous method is that u_h is allowed to be discontinuous across the boundary. Therefore, across the element, the following two different values are defined at the two sides of the boundary:

$$u_j^+ = \lim_{x \searrow x_j^+} u_h(x) \qquad u_j^- = \lim_{x \nearrow x_j^-} u_h(x) \tag{14.6}$$
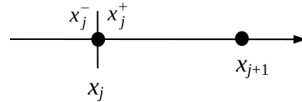
An illustration of the jump across x_j of element j: x_j and x_{j+1} mark the boundaries of the element

Conversely, we also have:

$$u_{j+1}^+ = \lim_{x \searrow x_{j+1}^+} u_h(x) \qquad u_{j+1}^- = \lim_{x \nearrow x_{j+1}^-} u_h(x) \tag{14.7}$$

It is key to remember that 1) u_h is discontinuous only at the element boundaries; 2) the solution u is smooth within (but excluding) the boundary. By this definition, the above equation contains the variables only within the integral limits of Ω_j . As a consequence, there is no direct coupling with other intervals or other elements. *The field values at a node, or the interface between two elements, are not unique.* They are calculated using the two limiting values approaching the interface from the two adjacent elements. This feature is certainly desirable for problems with internal discontinuities.

We can finally write for a single element:

$$u_{j+1}^- f_h(x_{j+1}) - u_j^+ f_h(x_j) - \int_{x_j}^{x_{j+1}} \left(u_h\frac{df_h}{dx} - gf_h(x) \right) dx = 0 \tag{14.8}$$

14.2 Steady state diffusion in 1D

Let us start simple with the 1D steady state heat conduction problem in 1D, given by the following equation:

$$\frac{d^2T}{dx^2} = 0 \qquad T(x = 0) = 0 \qquad T(x = 1) = 1 \qquad \text{on} \quad x \in [0, 1] \tag{14.9}$$

Although this equation is usually solved as is with its second-order derivative, it can also be written in a mixed form, using the heat flux q (a scalar in 1D):

$$
\begin{aligned}
-\frac{dq}{dx} &= 0 \\
q - \frac{dT}{dx} &= 0 \qquad x \in [0, 1]
\end{aligned}
\tag{14.10}
$$

and the boundary conditions remain unchanged.

We apply the standard approach to establish the weak forms of these two first-order ODEs, and we do so on an element e bound by nodes k and $k + 1$ with coordinates x_k and x_{k+1}

$$-\int_{x_k}^{x_{k+1}} \frac{dq}{dx}\tilde{f}(x)dx = -\left[q\tilde{f}\right]_{x_k}^{x_{k+1}} + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx}q(x)dx = 0 \tag{14.11}$$

$$\int_{x_k}^{x_{k+1}} \left(q - \frac{dT}{dx}\right)\overline{f}(x)dx = \int_{x_k}^{x_{k+1}} q(x)\overline{f}(x)dx - \left[T\overline{f}\right]_{x_k}^{x_{k+1}} + \int_{x_k}^{x_{k+1}} \frac{d\overline{f}}{dx}T(x)dx = 0 \tag{14.12}$$

where $\tilde{f}$ and $\overline{f}$ are test functions. We now must examine the term between square brackets. Inside the element, the test functions $\tilde{f}$ and $\overline{f}$ are well defined polynomials and we coin:

$$
\begin{aligned}
\tilde{f}_k^+ &= \tilde{f}(x_k^+) & (14.13) \\
\tilde{f}_{k+1}^- &= \tilde{f}(x_{k+1}^-) & (14.14) \\
\overline{f}_k^+ &= \overline{f}(x_k^+) & (14.15) \\
\overline{f}_{k+1}^- &= \overline{f}(x_{k+1}^-) & (14.16)
\end{aligned}
$$

Concerning q and T, we will for now give them values $\hat{q}_k$ and $\hat{T}_k$ at node k and $\hat{q}_{k+1}$ and $\hat{T}_{k+1}$ at node $k + 1$, and we will specify the hat quantities as follows:

$$
\hat{T}_k =
\begin{cases}
T_k^- & k = 1 \\
\frac{1}{2}(T_k^- + T_k^+) + \mathcal{C}(T_k^- - T_k^+) & k = 2, \dots N - 1 \\
T_k^+ & k = N
\end{cases}
$$

$$
\hat{q}_k =
\begin{cases}
q_k^+ - \mathcal{E}(T_k^- - T_k^+) & k = 1 \\
\frac{1}{2}(q_k^+ + q_k^-) - \mathcal{E}(T_k^- - T_k^+) - \mathcal{C}(q_k^- - q_k^+) & k = 2, \dots N - 1 \\
q_k^- - \mathcal{E}(T_k^- - T_k^+) & k = N
\end{cases}
\tag{14.17}
$$

where N is the number of nodes and where $\mathcal{C}$ and $\mathcal{E}$ are two constants.

Discuss the meaning/values of these!

Remark. *Note that $\hat{T}_k = T_1^-$ on the left boundary is consistent with $\hat{T}_k = \frac{1}{2}(T_k^- + T_k^+) + \mathcal{C}(T_k^- - T_k^+)$ provided $T_1^- = T_1^+$. The same goes for the right boundary, and the same reasoning applies for the heat flux terms $\hat{q}_k$.*

Inside an element bounded by nodes k and $k+1$, the temperature T and heat flux q are interpolated over an isoparametric linear element:

$$T_h(x) = \mathcal{N}_k(x)T_k^+ + \mathcal{N}_{k+1}(x)T_{k+1}^-$$

$$q_h(x) = \mathcal{N}_k(x)q_k^+ + \mathcal{N}_{k+1}(x)q_{k+1}^-$$

As in the (Continuous/Standard) Galerkin case of section 9.1, the test functions are taken to be the basis functions, and in this case for both temperature and flux.

There are four unknowns q_k^+, q_{k+1}^-, T_k^+ and T_{k+1}^- per element. All other q and T quantities in the above/following equations will need to find their way to the rhs.

- Eq. 14.11 becomes:

$$
\begin{aligned}
0 &= -\hat{q}_{k+1}\tilde{f}(x_{k+1}^-) + \hat{q}_k\tilde{f}(x_k^+) + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx}q_h(x)dx \\
&= -\hat{q}_{k+1}\tilde{f}_{k+1}^- + \hat{q}_k\tilde{f}_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx}(\mathcal{N}_k(x)q_k^+ + \mathcal{N}_{k+1}(x)q_{k+1}^-)dx \\
&= -\hat{q}_{k+1}\tilde{f}_{k+1}^- + \hat{q}_k\tilde{f}_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx}\mathcal{N}_k(x)dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx}\mathcal{N}_{k+1}(x)dx \cdot q_{k+1}^- \quad (14.18)
\end{aligned}
$$

  - We take $\tilde{f} = \mathcal{N}_k$ and by vertue of the properties of basis functions $\mathcal{N}$ we have:

$$
\begin{aligned}
\tilde{f}_k^+ &= \tilde{f}(x_k^+) = \mathcal{N}_k(x_k^+) = 1 \\
\tilde{f}_{k+1}^- &= \tilde{f}(x_{k+1}^-) = \mathcal{N}_k(x_{k+1}^+) = 0
\end{aligned}
$$

   so that

$$
\begin{aligned}
0 &= \hat{q}_k + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_k(x)dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_{k+1}(x)dx \cdot q_{k+1}^- \\
&= \frac{1}{2}(q_k^+ + q_k^-) - \mathcal{E}(T_k^- - T_k^+) - \mathcal{C}(q_k^- - q_k^+) \\
&\quad + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_k dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_{k+1}dx \cdot q_{k+1}^- \quad (14.19)
\end{aligned}
$$

  - We take $\tilde{f} = \mathcal{N}_{k+1}$ and likewise:

$$
\begin{aligned}
\tilde{f}_k^+ &= \tilde{f}(x_k^+) = \mathcal{N}_{k+1}(x_k^+) = 0 \\
\tilde{f}_{k+1}^- &= \tilde{f}(x_{k+1}^-) = \mathcal{N}_{k+1}(x_{k+1}^+) = 1
\end{aligned}
$$

   so that

$$
\begin{aligned}
0 &= -\hat{q}_{k+1} + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k(x)dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1}(x)dx \cdot q_{k+1}^- \\
&= -\left[\frac{1}{2}(q_{k+1}^+ + q_{k+1}^-) - \mathcal{E}(T_{k+1}^- - T_{k+1}^+) - \mathcal{C}(q_{k+1}^- - q_{k+1}^+)\right] \\
&\quad + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1}dx \cdot q_{k+1}^- \quad (14.20)
\end{aligned}
$$

   and finally

$$
\begin{aligned}
&\int_{x_k}^{x_{k+1}} \begin{pmatrix} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_k}{dx}\mathcal{N}_{k+1} \\ \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1} \end{pmatrix} dx \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\mathcal{C} + \frac{1}{2})q_k^+ \\ (\mathcal{C} - \frac{1}{2})q_{k+1}^- \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^+ \\ \mathcal{E}T_{k+1}^- \end{pmatrix} \\
&= \begin{pmatrix} (\mathcal{C} - \frac{1}{2})q_k^- \\ (\mathcal{C} + \frac{1}{2})q_{k+1}^+ \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^- \\ \mathcal{E}T_{k+1}^+ \end{pmatrix} \quad (14.21)
\end{aligned}
$$

590

- Eq. 14.12 becomes:

$$
\begin{aligned}
0 &= -[T\overline{f}]_{x_k}^{x_{k+1}} + \int_{x_k}^{x_{k+1}} q_h(x)\overline{f}(x)dx + \int_{x_k}^{x_{k+1}} \frac{d\overline{f}}{dx}T_h(x)dx \\
&= -\hat{T}_{k+1}\overline{f}_{k+1}^- + \hat{T}_k\overline{f}_k^+ + \int_{x_k}^{x_{k+1}} q_h(x)\overline{f}(x)dx + \int_{x_k}^{x_{k+1}} \frac{d\overline{f}}{dx}T_h(x)dx
\end{aligned}
$$

  – We take $\overline{f} = \mathcal{N}_k$:

$$
\begin{aligned}
\overline{f}_k^+ &= \overline{f}(x_k^+) = \mathcal{N}_k(x_k^+) = 1 \\
\overline{f}_{k+1}^- &= \overline{f}(x_{k+1}^-) = \mathcal{N}_k(x_{k+1}^-) = 0
\end{aligned}
$$

  so that

$$
\begin{aligned}
0 &= \hat{T}_k + \int_{x_k}^{x_{k+1}} q_h(x)\mathcal{N}_k dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}T_h(x)dx \\
&= \frac{1}{2}(T_k^- + T_k^+) + \mathcal{C}(T_k^- - T_k^+) \\
&+ \int_{x_k}^{x_{k+1}} (\mathcal{N}_k(x)q_k^+ + \mathcal{N}_{k+1}(x)q_{k+1}^-)\mathcal{N}_k dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx}(\mathcal{N}_k(x)T_k^+ + \mathcal{N}_{k+1}(x)T_{k+1}^-)dx
\end{aligned}
\tag{14.22}
$$

  – We take $\overline{f} = \mathcal{N}_{k+1}$:

$$
\begin{aligned}
\overline{f}_k^+ &= \overline{f}(x_k^+) = \mathcal{N}_{k+1}(x_k^+) = 0 \\
\overline{f}_{k+1}^- &= \overline{f}(x_{k+1}^-) = \mathcal{N}_{k+1}(x_{k+1}^-) = 1
\end{aligned}
$$

  so that

$$
\begin{aligned}
0 &= -\hat{T}_{k+1} + \int_{x_k}^{x_{k+1}} q_h(x)\mathcal{N}_{k+1}dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}T_h(x)dx \\
&= -\left[\frac{1}{2}(T_{k+1}^- + T_{k+1}^+) + \mathcal{C}(T_{k+1}^- - T_{k+1}^+)\right] \tag{14.23}\\
&+ \int_{x_k}^{x_{k+1}} (\mathcal{N}_k(x)q_k^+ + \mathcal{N}_{k+1}(x)q_{k+1}^-)\mathcal{N}_{k+1}dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}(\mathcal{N}_k(x)T_k^+ + \mathcal{N}_{k+1}(x)T_{k+1}^-)dx
\end{aligned}
\tag{14.24}
$$

and finally

$$
\begin{aligned}
&\int_{x_k}^{x_{k+1}} \begin{pmatrix} \mathcal{N}_k\mathcal{N}_k & \mathcal{N}_k\mathcal{N}_{k+1} \\ \mathcal{N}_{k+1}\mathcal{N}_k & \mathcal{N}_{k+1}\mathcal{N}_{k+1} \end{pmatrix} dx \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \int_{x_k}^{x_{k+1}} \begin{pmatrix} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_k}{dx}\mathcal{N}_{k+1} \\ \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1} \end{pmatrix} dx \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} \\
&+ \begin{pmatrix} (\frac{1}{2} - \mathcal{C})T_k^+ \\ -(\mathcal{C} + \frac{1}{2})T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C} + \frac{1}{2})T_k^- \\ (\frac{1}{2} - \mathcal{C})T_{k+1}^+ \end{pmatrix}
\end{aligned}
\tag{14.25}
$$

We will also use the results obtained in Appendix C for 1D linear elements:

$$
\boldsymbol{M}^e = \int_{\Omega_e} \vec{\mathcal{N}}^T \vec{\mathcal{N}} dV = \int_{\Omega_e} \begin{pmatrix} \mathcal{N}_k \mathcal{N}_k & \mathcal{N}_k \mathcal{N}_{k+1} \\ \mathcal{N}_{k+1} \mathcal{N}_k & \mathcal{N}_{k+1} \mathcal{N}_{k+1} \end{pmatrix} dV = \frac{h}{2} \frac{1}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}
$$

$$
\boldsymbol{K}^e = \int_{\Omega_e} \begin{pmatrix} \frac{d\mathcal{N}_k}{dx} \mathcal{N}_k & \frac{d\mathcal{N}_k}{dx} \mathcal{N}_{k+1} \\ \frac{d\mathcal{N}_{k+1}}{dx} \mathcal{N}_k & \frac{d\mathcal{N}_{k+1}}{dx} \mathcal{N}_{k+1} \end{pmatrix} dV = \frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \tag{14.26}
$$

Filling this into equations (14.21) and (14.25), gives

$$
\boldsymbol{K}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\mathcal{C}+\frac{1}{2})q_k^+ \\ (\mathcal{C}-\frac{1}{2})q_{k+1}^- \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^+ \\ \mathcal{E}T_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C}-\frac{1}{2})q_k^- \\ (\mathcal{C}+\frac{1}{2})q_{k+1}^+ \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^- \\ \mathcal{E}T_{k+1}^+ \end{pmatrix}
$$

$$
\boldsymbol{M}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \boldsymbol{K}^e \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\frac{1}{2}-\mathcal{C})T_k^+ \\ -(\mathcal{C}+\frac{1}{2})T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C}+\frac{1}{2})T_k^- \\ (\frac{1}{2}-\mathcal{C})T_{k+1}^+ \end{pmatrix} \tag{14.27}
$$

which becomes

$$
\begin{pmatrix} \mathcal{C} & -\frac{1}{2} \\ \frac{1}{2} & \mathcal{C} \end{pmatrix} \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} \mathcal{E} & 0 \\ 0 & \mathcal{E} \end{pmatrix} \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C}-\frac{1}{2})q_k^- + \mathcal{E}T_k^- \\ (\mathcal{C}+\frac{1}{2})q_{k+1}^+ + \mathcal{E}T_{k+1}^+ \end{pmatrix}
$$

$$
\frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} -\mathcal{C} & -\frac{1}{2} \\ \frac{1}{2} & -\mathcal{C} \end{pmatrix} \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\frac{1}{2}+\mathcal{C})T_k^- \\ (\frac{1}{2}-\mathcal{C})T_{k+1}^+ \end{pmatrix}
$$

Combining these equations gives the expression for the linear system for the element under consideration:

$$
\begin{pmatrix} \frac{h}{3} & \frac{h}{6} & -\mathcal{C} & -\frac{1}{2} \\ \frac{h}{6} & \frac{h}{3} & \frac{1}{2} & -\mathcal{C} \\ \mathcal{C} & -\frac{1}{2} & \mathcal{E} & 0 \\ \frac{1}{2} & \mathcal{C} & 0 & \mathcal{E} \end{pmatrix} \begin{pmatrix} q_k^+ \\ q_{k+1}^- \\ T_k^+ \\ T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\frac{1}{2}+\mathcal{C})T_k^- \\ (\frac{1}{2}-\mathcal{C})T_{k+1}^+ \\ -(\frac{1}{2}-\mathcal{C})q_k^- + \mathcal{E}T_k^- \\ (\frac{1}{2}+\mathcal{C})q_{k+1}^+ + \mathcal{E}T_{k+1}^+ \end{pmatrix} \tag{14.28}
$$

Left boundary Special care must be taken for the two elements on the boundaries of the domain. On the left, we have

$$
\hat{q}_1 = q_1^+ - \mathcal{E}(T_1^- - T_1^+)
$$
$$
\hat{T}_1 = T_1^-
$$

Eq. (14.19) becomes:

$$
0 = \hat{q}_k + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} \mathcal{N}_k(x) dx \cdot q_k^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} \mathcal{N}_{k+1}(x) dx \cdot q_{k+1}^-
$$

$$
= q_1^+ - \mathcal{E}(T_k^- - T_1^+) + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} \mathcal{N}_k dx \cdot q_1^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} \mathcal{N}_{k+1} dx \cdot q_2^- \tag{14.29}
$$

Eq. (14.22) becomes:

$$
0 = \hat{T}_1 + \int_{x_k}^{x_{k+1}} q_h(x) \mathcal{N}_k dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} T_h(x) dx
$$

$$
= T_1^- + \int_{x_k}^{x_{k+1}} (\mathcal{N}_k(x)q_1^+ + \mathcal{N}_{k+1}(x)q_2^-) \mathcal{N}_k dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_k}{dx} (\mathcal{N}_k(x)T_1^+ + \mathcal{N}_{k+1}(x)T_2^-) dx \tag{14.30}
$$

Eq. (14.28) then becomes:

$$
\begin{pmatrix} \frac{h}{3} & \frac{h}{6} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{h}{6} & \frac{h}{3} & \frac{1}{2} & -\mathcal{C} \\ \frac{1}{2} & -\frac{1}{2} & \mathcal{E} & 0 \\ \frac{1}{2} & \mathcal{C} & 0 & \mathcal{E} \end{pmatrix} \begin{pmatrix} q_1^+ \\ q_2^- \\ T_1^+ \\ T_2^- \end{pmatrix} = \begin{pmatrix} -T_1^- \\ (\frac{1}{2}-\mathcal{C})T_{k+1}^+ \\ \mathcal{E}T_k^- \\ (\frac{1}{2}+\mathcal{C})q_{k+1}^+ + \mathcal{E}T_{k+1}^+ \end{pmatrix} \tag{14.31}
$$

Right boundary The element is composed of nodes $N-1$ and N. The fluxes are

$$
\begin{aligned}
\hat{q}_N &= q_N^+ - \mathcal{E}(T_N^- - T_N^+) \\
\hat{T}_N &= T_N^-
\end{aligned}
$$

Eq. (14.20) becomes:

$$
\begin{aligned}
0 &= -\hat{q}_N + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k(x)dx \cdot q_{N-1}^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1}(x)dx \cdot q_N^- \\
&= -\left[q_N^+ - \mathcal{E}(T_N^- - T_N^+)\right] + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k dx \cdot q_{N-1}^+ + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1} dx \cdot q_N^- \quad (14.32)
\end{aligned}
$$

Eq. (14.24) becomes:

$$
\begin{aligned}
0 &= -\hat{T}_N + \int_{x_k}^{x_{k+1}} q_h(x)\mathcal{N}_{k+1}dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}T_h(x)dx \\
&= -T_N^- + \int_{x_k}^{x_{k+1}} (\mathcal{N}_k(x)q_{N-1}^+ + \mathcal{N}_{k+1}(x)q_N^-)\mathcal{N}_{k+1}dx + \int_{x_k}^{x_{k+1}} \frac{d\mathcal{N}_{k+1}}{dx}(\mathcal{N}_k(x)T_{N-1}^+ + \mathcal{N}_{k+1}(x)T_N^-)dx
\end{aligned}
$$

Eq. (14.28) then becomes:

$$
\begin{pmatrix}
\frac{h}{3} & \frac{h}{6} & -\mathcal{C} & -\frac{1}{2} \\
\frac{h}{6} & \frac{h}{3} & \frac{1}{2} & \frac{1}{2} \\
\mathcal{C} & -\frac{1}{2} & \mathcal{E} & 0 \\
\frac{1}{2} & -\frac{1}{2} & 0 & \mathcal{E}
\end{pmatrix}
\begin{pmatrix}
q_{N-1}^+ \\
q_N^- \\
T_{N-1}^+ \\
T_N^-
\end{pmatrix}
=
\begin{pmatrix}
-(\frac{1}{2}+\mathcal{C})T_{N-1}^- \\
T_N^+ \\
-(\frac{1}{2}-\mathcal{C})q_{N-1}^- + \mathcal{E}T_{N-1}^- \\
\mathcal{E}T_N^+
\end{pmatrix}
\quad (14.33)
$$

Solving strategies Following Li [825], there are three main strategies:

- Successive substitution: all the variables are initialized to zero. Eq. (14.31) is solved to obtain the data for the first element, where boundary conditions are specified. Then Eq. (14.28) is used for all interior elements. Finally Eq. (14.33) is used for the last element. This procedure is carried out until all fields have converged.

- Global assembly: this approach is identical to the one we have taken so far with the continuous Galerkin Finite Element method. We form a large global matrix and then solve the linear system to obtain the solution. The disadvantage of this approach lies in the size of the generated marix: each node counts 4 dofs so the assembled matrix will be about 4 times as big as in the standard FE case.

- Elimination then asssembly: one can first eliminate the variable q and solve for the temperature T only. This speedis up the calculations, but also increases the bandwidth of the element matrix. Li [825] states: "Further comparison shows that the saving in CPU time for solving T alone is less significant than the $q-T$ iterative solution, in particular, for 3-D problems."

  Eq. (14.28) can be rewritten:

$$
\begin{pmatrix}
\boldsymbol{M}_e & \boldsymbol{C}_1 \\
\boldsymbol{C}_2 & \boldsymbol{E}
\end{pmatrix}
\cdot
\begin{pmatrix}
\vec{q} \\
\vec{T}
\end{pmatrix}
=
\begin{pmatrix}
\vec{f} \\
\vec{g}
\end{pmatrix}
\quad (14.34)
$$

  The unknown of the original ODE is temperature so this is the quantity we are after. The first line of the matrix can be written

$$
\boldsymbol{M}_e \cdot \vec{q} + \boldsymbol{C}_1 \cdot \vec{T} = \vec{f}
$$

or,

$$\vec{q} = \boldsymbol{M}_e^{-1} \cdot (\vec{f} - \boldsymbol{C}_1 \cdot \vec{T})$$

The second line of the matrix is

$$\boldsymbol{C}_2 \cdot \vec{q} + \boldsymbol{E} \cdot \vec{T} = \vec{g}$$

and we then replace $\vec{q}$ by the expression above:

$$\boldsymbol{C}_2 \cdot [\boldsymbol{M}_e^{-1} \cdot (\vec{f} - \boldsymbol{C}_1 \cdot \vec{T})] + \boldsymbol{E} \cdot \vec{T} = \vec{g}$$

or,

$$-\boldsymbol{C}_2 \cdot \boldsymbol{M}_e^{-1} \cdot \boldsymbol{C}_1 \cdot \vec{T} + \boldsymbol{E} \cdot \vec{T} = \vec{g} - \boldsymbol{C}_2 \cdot \boldsymbol{M}_e^{-1} \cdot \vec{f}$$

and finally

$$[\boldsymbol{E} - \boldsymbol{C}_2 \cdot \boldsymbol{M}_e^{-1} \cdot \boldsymbol{C}_1] \cdot \vec{T} = \vec{g} - \boldsymbol{C}_2 \cdot \boldsymbol{M}_e^{-1} \cdot \vec{f}$$

Note that the matrix will still be twice as big than in the standard FEM case since each node counts two temperature dofs.

Choice of $\mathcal{C}$ and $\mathcal{E}$ Li [825] shows satisfying results for $\mathcal{E} = 4$ and $\mathcal{C} = -1/2, 0, 1/2$ or $\mathcal{E} = 0$ and $\mathcal{C} = 1, 4, 10$.

Remark. *Aside from the shear complexity of the above derivations as compared to those for the SG method, the fact that we have two tuning parameters $\mathcal{E}$ and $\mathcal{C}$ is a real drawback.*

14.3   Time-dependent diffusion PDE in 1D

Starting from the simple transient 1-D heat conduction problem similar to the steady state heat conduction problem only with added time dependence:

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} \qquad T(x = 0) = 0 \qquad T(x = 1) = 1 \qquad \text{on} \quad x \in [0, 1] \tag{14.35}$$

Once again we split this system into two seperate first order equations:

$$\begin{aligned}
\frac{\partial T}{\partial t} - \frac{\partial q}{\partial x} &= 0 \\
\frac{\partial T}{\partial x} - q &= 0
\end{aligned} \tag{14.36}$$

We apply the standard approach to establish the weak forms of these two first-order PDEs, and we do so on an element e bound by nodes k and $k + 1$ with coordinates x_k and x_{k+1}

$$-\int_{x_k}^{x_{k+1}} \left(\frac{\partial T}{\partial t} - \frac{dq}{dx} \right) \tilde{f}(x) dx = \int_{x_k}^{x_{k+1}} \frac{\partial T}{\partial t} \tilde{f}(x) dx - \left[q\tilde{f} \right]_{x_k}^{x_{k+1}} + \int_{x_k}^{x_{k+1}} \frac{d\tilde{f}}{dx} q(x) dx = 0 \tag{14.37}$$

$$\int_{x_k}^{x_{k+1}} \left(q - \frac{dT}{dx} \right) \overline{f}(x) dx = \int_{x_k}^{x_{k+1}} q(x)\overline{f}(x) dx - \left[T\overline{f} \right]_{x_k}^{x_{k+1}} + \int_{x_k}^{x_{k+1}} \frac{d\overline{f}}{dx} T(x) dx = 0 \tag{14.38}$$

where $\tilde{f}$ and $\overline{f}$ are test functions.

In what follows we coin $\dot{T} = \partial T / \partial t$ (for convenience of notation). We once again recover Equations (14.21) and (14.25), although with an additional time derivative term.

Filling this into equations (14.21) and (14.25), gives

$$\boldsymbol{K}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \boldsymbol{M}^e \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\mathcal{C} + \frac{1}{2})q_k^+ \\ (\mathcal{C} - \frac{1}{2})q_{k+1}^- \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^+ \\ \mathcal{E}T_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C} - \frac{1}{2})q_k^- \\ (\mathcal{C} + \frac{1}{2})q_{k+1}^+ \end{pmatrix} + \begin{pmatrix} \mathcal{E}T_k^- \\ \mathcal{E}T_{k+1}^+ \end{pmatrix}$$

$$\boldsymbol{M}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \boldsymbol{K}^e \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\frac{1}{2} - \mathcal{C})T_k^+ \\ -(\mathcal{C} + \frac{1}{2})T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C} + \frac{1}{2})T_k^- \\ (\frac{1}{2} - \mathcal{C})T_{k+1}^+ \end{pmatrix} \qquad (14.39)$$

In what follows we set $\mathcal{E} = 0$ so that we have

$$\boldsymbol{K}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \boldsymbol{M}^e \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\mathcal{C} + \frac{1}{2})q_k^+ \\ (\mathcal{C} - \frac{1}{2})q_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C} - \frac{1}{2})q_k^- \\ (\mathcal{C} + \frac{1}{2})q_{k+1}^+ \end{pmatrix}$$

$$\boldsymbol{M}^e \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \boldsymbol{K}^e \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\frac{1}{2} - \mathcal{C})T_k^+ \\ -(\mathcal{C} + \frac{1}{2})T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C} + \frac{1}{2})T_k^- \\ (\frac{1}{2} - \mathcal{C})T_{k+1}^+ \end{pmatrix} \qquad (14.40)$$

Using the expressions for $\boldsymbol{M}^e$ and $\boldsymbol{K}^e$ obtained in Appendix C for 1D linear elements we arrive at

$$\frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\mathcal{C} + \frac{1}{2})q_k^+ \\ (\mathcal{C} - \frac{1}{2})q_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C} - \frac{1}{2})q_k^- \\ (\mathcal{C} + \frac{1}{2})q_{k+1}^+ \end{pmatrix}$$

$$\frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + \begin{pmatrix} (\frac{1}{2} - \mathcal{C})T_k^+ \\ -(\mathcal{C} + \frac{1}{2})T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C} + \frac{1}{2})T_k^- \\ (\frac{1}{2} - \mathcal{C})T_{k+1}^+ \end{pmatrix} \qquad (14.41)$$

which simplifies to

$$\begin{pmatrix} C & -1/2 \\ 1/2 & C \end{pmatrix} \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} h/3 & h/6 \\ h/6 & h/3 \end{pmatrix} \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} = \begin{pmatrix} (\mathcal{C} - \frac{1}{2})q_k^- \\ (\mathcal{C} + \frac{1}{2})q_{k+1}^+ \end{pmatrix}$$

$$\begin{pmatrix} h/3 & h/6 \\ h/6 & h/3 \end{pmatrix} \cdot \begin{pmatrix} q_k^+ \\ q_{k+1}^- \end{pmatrix} + \begin{pmatrix} -C & -1/2 \\ 1/2 & -C \end{pmatrix} \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} = \begin{pmatrix} -(\mathcal{C} + \frac{1}{2})T_k^- \\ (\frac{1}{2} - \mathcal{C})T_{k+1}^+ \end{pmatrix} \qquad (14.42)$$

or,

$$\boldsymbol{C}_1 \vec{q} + \boldsymbol{M} \vec{T} = \vec{f}$$
$$\boldsymbol{M} \vec{q} + \boldsymbol{C}_2 \vec{T} = \vec{g}$$

so

$$\vec{q} = \boldsymbol{M}^{-1}(\vec{g} - \boldsymbol{C}_2 \vec{T})$$

and then

$$\boldsymbol{C}_1 [\boldsymbol{M}^{-1}(\vec{g} - \boldsymbol{C}_2 \vec{T})] + \boldsymbol{M} \vec{T} = \vec{f}$$

NOT REALLY FINISHED...

14.4 Time-dependent advection PDE in 1D

Starting from the 1-D advection equation:

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} = 0 \tag{14.43}$$

where T is the temperature and u the velocity. As shown before we start by discretizing the domain into a collection of elements. Then the above equation can be integrated over the element which is bounded by nodes x_k and x_{k+1}.

$$\int_{x_k}^{x_{k+1}} \left(\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x}\right) \tilde{f}(x)dx = \int_{x_k}^{x_{k+1}} \tilde{f}(x)\frac{\partial T}{\partial t}dx + \left[uT\tilde{f}\right]_{x_k}^{x_{k+1}} - \int_{x_k}^{x_{k+1}} \frac{\partial \tilde{f}}{\partial x}uTdx = 0$$

with $\tilde{f}$ the test function. Inside the elements the test functions are defined by well defined polynomials. We once again define

$$\begin{aligned} \tilde{f}_k^+ &= \tilde{f}(x_k^+) \\ \tilde{f}_{k+1}^- &= \tilde{f}(x_{k+1}^-) \end{aligned}$$

$$\int_{x_k}^{x_{k+1}} \left(\tilde{f}(x)\frac{\partial T_h}{\partial t} - \frac{\partial \tilde{f}}{\partial x}uT_h\right)dx + \tilde{f}(x_{k+1})\widehat{uT}(T_{k+1}^-, T_{k+1}^+) - \tilde{f}(x_k)\widehat{uT}(T_k^-, T_k^+) = 0 \tag{14.44}$$

For a constant u or a linear problem, an effective numerical flux is the Lax-Friedrichs flux:

$$\widehat{uT}(a,b) = u\frac{(a+b)}{2} - |u|\frac{(b-a)}{2} \tag{14.45}$$

when $u > 0$ this flux then simply becomes:

$$uT(a,b) = ua \tag{14.46}$$

which is in essence an upwinding scheme. Filling this into equation (14.44) gives[2]:

$$\int_{x_k}^{x_{k+1}} \left(\tilde{f}(x)\frac{\partial T_h}{\partial t} - \frac{\partial \tilde{f}}{\partial x}uT_h\right)dx + \tilde{f}_{k+1}^- uT_{k+1}^- - \tilde{f}_k^- uT_k^- = 0 \tag{14.47}$$

The function T_h inside the element can be approximated as follows:

$$T_h(x) = \sum_{i=1}^{m} \mathcal{N}_i(x)T_i = \mathcal{N}_k(x)T_k^+ + \mathcal{N}_{k+1}(x)T_{k+1}^- \tag{14.48}$$

where the red quantities are the unknown dofs. In what follows we coin $\dot{T} = \partial T/\partial t$ so

$$\dot{T}_h(x) = \sum_{i=1}^{m} \mathcal{N}_i(x)\dot{T}_i = \mathcal{N}_k(x)\dot{T}_k^+ + \mathcal{N}_{k+1}(x)\dot{T}_{k+1}^- \tag{14.49}$$

Taking $\tilde{f}(x) = \mathcal{N}_k(x)$ and then $\tilde{f}(x) = \mathcal{N}_{k+1}(x)$ we arrive at

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_k(x)\dot{T}_h - \frac{\partial \mathcal{N}_k}{\partial x}uT_h\right)dx + \underbrace{\mathcal{N}_k(x_{k+1}^-)}_{=0} uT_{k+1}^- - \underbrace{\mathcal{N}_k(x_k^-)}_{=1} uT_k^- = 0 \tag{14.50}$$

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_{k+1}(x)\dot{T}_h - \frac{\partial \mathcal{N}_{k+1}}{\partial x}uT_h\right)dx + \underbrace{\mathcal{N}_{k+1}(x_{k+1}^-)}_{=1} uT_{k+1}^- - \underbrace{\mathcal{N}_{k+1}(x_k^-)}_{=0} uT_k^- = 0. \tag{14.51}$$

---

[2]why do we only keep T^-? because of the velocity direction?

The underbraced terms are either 0 or 1 due to the properties of the basis functions $\mathcal{N}_i(\vec{r}_j) = \delta_{ij}$. Finally:

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_k(x)\dot{T}_h - \frac{\partial \mathcal{N}_k}{\partial x} u T_h \right) dx - u T_k^- = 0 \qquad (14.52)$$

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_{k+1}(x)\dot{T}_h - \frac{\partial \mathcal{N}_{k+1}}{\partial x} u T_h \right) dx + u T_{k+1}^- = 0 \qquad (14.53)$$

We now use Eqs. (14.48) and (14.49) in Eq. (14.52) and Eq. (14.53):

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_k[\mathcal{N}_k\dot{T}_k^+ + \mathcal{N}_{k+1}\dot{T}_{k+1}^-] - \frac{\partial \mathcal{N}_k}{\partial x} u[\mathcal{N}_k T_k^+ + \mathcal{N}_{k+1} T_{k+1}^-] \right) dx - u T_k^- = 0 \quad (14.54)$$

$$\int_{x_k}^{x_{k+1}} \left(\mathcal{N}_{k+1}[\mathcal{N}_k\dot{T}_k^+ + \mathcal{N}_{k+1}\dot{T}_{k+1}^-] - \frac{\partial \mathcal{N}_{k+1}}{\partial x} u[\mathcal{N}_k T_k^+ + \mathcal{N}_{k+1} T_{k+1}^-] \right) dx + u T_{k+1}^- = 0 \quad (14.55)$$

Defining again (see Appendix C)

$$\boldsymbol{M}_e = \int_{x_k}^{x_{k+1}} \begin{pmatrix} \mathcal{N}_k\mathcal{N}_k & \mathcal{N}_k\mathcal{N}_{k+1} \\ \mathcal{N}_{k+1}\mathcal{N}_k & \mathcal{N}_{k+1}\mathcal{N}_{k+1} \end{pmatrix} dx = \frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

and

$$\boldsymbol{K}^e = \int_{\Omega_e} \begin{pmatrix} \frac{d\mathcal{N}_k}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_k}{dx}\mathcal{N}_{k+1} \\ \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_k & \frac{d\mathcal{N}_{k+1}}{dx}\mathcal{N}_{k+1} \end{pmatrix} dV = \frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$$

This results in:

$$\boldsymbol{M}_e \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} - u\boldsymbol{K}_e \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + u \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ T_{k+1}^- \end{pmatrix} - \begin{pmatrix} u T_k^- \\ 0 \end{pmatrix} = 0 \qquad (14.56)$$

or,

$$\boldsymbol{M}_e \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} = u \left[\boldsymbol{K}_e - \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right] \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + u \begin{pmatrix} T_k^- \\ 0 \end{pmatrix} \qquad (14.57)$$

$$\boldsymbol{M}_e \cdot \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} = u\frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + u \begin{pmatrix} T_k^- \\ 0 \end{pmatrix} \qquad (14.58)$$

$$\begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} = u\boldsymbol{M}_e^{-1} \cdot \frac{1}{2} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + u\boldsymbol{M}_e^{-1} \cdot \begin{pmatrix} T_k^- \\ 0 \end{pmatrix} \qquad (14.59)$$

We have already established that

$$\boldsymbol{M}_e^{-1} = \frac{2}{h} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

so

$$\boxed{ \begin{pmatrix} \dot{T}_k^+ \\ \dot{T}_{k+1}^- \end{pmatrix} = \frac{u}{h} \begin{pmatrix} -3 & -1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} T_k^+ \\ T_{k+1}^- \end{pmatrix} + \frac{u}{h} \begin{pmatrix} 4 \\ -2 \end{pmatrix} T_k^- } \qquad (14.60)$$

Using the same first order Runge-Kutta method as in the previous section,

$$T^k(t + \delta t) = T_k(t) + \delta t\, \dot{T}_k \qquad (14.61)$$

we multiply the equation by δt and we obtain

$$\begin{pmatrix} T_k^+(t + \delta t) \\ T_{k+1}^-(t + \delta t) \end{pmatrix} = \begin{pmatrix} T_k^+(t) \\ T_{k+1}^-(t) \end{pmatrix} + \frac{u\delta t}{h} \begin{pmatrix} -3 & -1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} T_k^+(t) \\ T_{k+1}^-(t) \end{pmatrix} + \frac{u\delta t}{h} \begin{pmatrix} 4 \\ -2 \end{pmatrix} T_k^-(t + \delta t) \quad (14.62)$$

and finally

$$\begin{pmatrix} T_k^+(t+\delta t) \\ T_{k+1}^-(t+\delta t) \end{pmatrix} = \left[\mathbf{1} + \frac{u\delta t}{h} \begin{pmatrix} -3 & -1 \\ 3 & -1 \end{pmatrix} \right] \cdot \begin{pmatrix} T_k^+(t) \\ T_{k+1}^-(t) \end{pmatrix} + \frac{u\delta t}{h} \begin{pmatrix} 4 \\ -2 \end{pmatrix} T_k^-(t+\delta t) \quad (14.63)$$

Also, since $C = u\delta t/h$, then the equation above can also be written

$$\begin{pmatrix} T_k^+(t+\delta t) \\ T_{k+1}^-(t+\delta t) \end{pmatrix} = \left[\mathbf{1} + C \begin{pmatrix} -3 & -1 \\ 3 & -1 \end{pmatrix} \right] \cdot \begin{pmatrix} T_k^+(t) \\ T_{k+1}^-(t) \end{pmatrix} + C \begin{pmatrix} 4 \\ -2 \end{pmatrix} T_k^-(t+\delta t) \quad (14.64)$$

This problem can be solved starting from the left boundary and sweeping through all the elements. The updated values of adjacent elements are used in the calculation of the next element as soon as this element becomes available which is why the last term T_k^- is taken at $t + \delta t$.

14.5 Steady-state diffusion in 2D

Let us start from the 2D steady state heat diffusion equation:

$$\vec{\nabla} \cdot k\vec{\nabla}T + H = 0 \tag{14.65}$$

Just as in the 1D case this equation can be split in two separate first order differential equations:

$$\underbrace{-\vec{\nabla} \cdot \vec{q} + H = 0}_{\text{ODE 1}} \quad ; \quad \underbrace{\vec{q} = -k\vec{\nabla}T}_{\text{ODE 2}} \tag{14.66}$$

Let N_i^θ be the temperature basis functions so that the temperature inside an element is given by

$$T_h(\vec{r}) = \sum_{i=1}^{m} N_i^\theta(\vec{r})\, T_i = \vec{N^\theta} \cdot \vec{T} \tag{14.67}$$

where $\vec{T}$ is a vector of length m, the number of nodes per element. Similarly we let the basis function for the heat flux be such that

$$q_h(\vec{r}) = \sum_{i=1}^{m} N_i^q(\vec{r}) q_i = \vec{N^q} \cdot \vec{q} \tag{14.68}$$

$$q_h(\vec{r}) = \sum_{i=1}^{m} N_i^q(\vec{r}) q_i = \vec{N^q} \cdot \vec{q} \tag{14.69}$$

where $\vec{q}_x$, $\vec{q}_y$ and $\vec{N^q}$ are vectors of length m too. Implicitly if m is the same for temperature and heat flux, then $N^\theta = N^q$. Let us establish the weak forms of the 1st order ODEs.

ODE 1 This results in:

$$\int_\Omega N_i^\theta \vec{\nabla} \cdot \vec{q}\, dV = \int_\Omega N_i^\theta H\, dV \tag{14.70}$$

Using the product rule $\vec{\nabla} \cdot (N_i^\theta \vec{q}) = N_i^\theta \vec{\nabla} \cdot \vec{q} + \vec{\nabla} N_i^\theta \cdot \vec{q}$ then we can write

$$N_i^\theta \vec{\nabla} \cdot \vec{q} = \vec{\nabla} \cdot (N_i^\theta \vec{q}) - \vec{\nabla} N_i^\theta \cdot \vec{q}$$

which we insert in Eq. (14.70) results in

$$\int_\Omega \vec{\nabla} \cdot (N_i^\theta \vec{q})\, dV - \int_\Omega \vec{\nabla} N_i^\theta \cdot \vec{q}\, dV = \int_\Omega N_i^\theta H\, dV \tag{14.71}$$

Using the divergence theorem $\int_\Omega \vec{\nabla} \cdot \vec{F}\, d\Omega = \int_{d\Omega} (\vec{F} \cdot \vec{n})\, dS$ applied to Eq. (14.71) leads to

$$\int_{d\Omega} N_i^\theta \vec{\hat{q}} \cdot \vec{n}\, dS - \int_\Omega \vec{\nabla} N_i^\theta \cdot \vec{q}\, dV = \int_\Omega N_i^\theta H\, dV \tag{14.72}$$

Here $\vec{n}$ is the outward vector everywhere on the boundary. The exact solution $\vec{q} = (q_x, q_y)$ can be approximated with $\vec{q}_h$ in a finite element space, same for the approximation of the flux at the boundary $\hat{q} = \hat{q}_h$ which takes a special form in the context of the DG methods (as reflected by the presence of the hat). We decompose the resulting equation in 3 terms A, B and C:

$$\underbrace{\int_{d\Omega} N_i^\theta \vec{\hat{q}_h} \cdot \vec{n}\, dS}_{A} - \underbrace{\int_\Omega \vec{\nabla} N_i^\theta \cdot \vec{q_h}\, dV}_{B} = \underbrace{\int_\Omega N_i^\theta H\, dV}_{C} \tag{14.73}$$

Note that terms B and C are explained in Section 9.4. In all what follows blue symbols belong the the element under consideration and brown symbols belong to its neighbour(s).
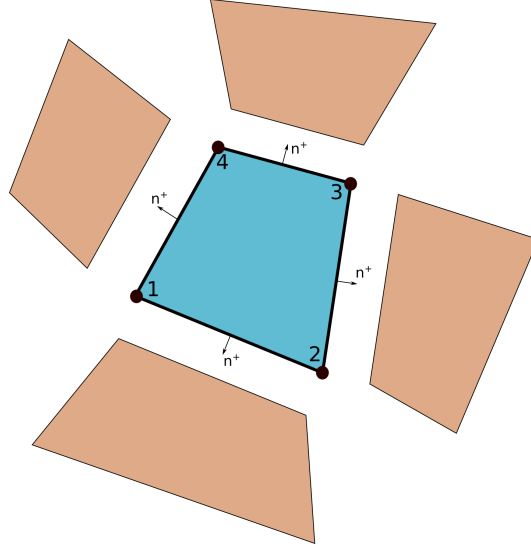
$\vec{n}^+$ indicates the outward vector at the boundary and $\vec{n}^-$ is the outward vector of the neighbouring element.

Let us turn to Li's book for useful definitions:

- Definition of jump operators The square brackets denote the jump operator:

$$
\begin{aligned}
[\vec{q}_h] &= \vec{q}_h \cdot \vec{n}^+ + \vec{q}_h \cdot \vec{n}^- \quad \text{or} \quad [\vec{q}_h] = (\vec{q}_h - \vec{q}_h) \cdot \vec{n}^+ \\
[T_h] &= T_h \vec{n}^+ + T_h \vec{n}^- \quad \text{or} \quad [T_h] = (T_h - T_h)\, \vec{n}^+
\end{aligned}
\tag{14.74}
$$

Note that $[\vec{q}_h]$ is a scalar function which involves the normal components only; while $[T_h]$ is a vector function.

- Definition of average operators The curly brackets indicate the average operator

$$
\begin{aligned}
\{\vec{q}_h\} &= \frac{1}{2}(\vec{q}_h + \vec{q}_h) \quad \text{so} \quad \{q_h\}_x = \frac{1}{2}(q_h\!\!_x + q_h\!\!_x) \quad \text{and} \quad \{q_h\}_y = \frac{1}{2}(q_h\!\!_y + q_h\!\!_y) \\
\{T_h\} &= \frac{1}{2}(T_h + T_h)
\end{aligned}
$$

Note that $\{\vec{q}_h\}$ is a vector, while $\{T_h\}$ is a scalar.

- Definitions of fluxes In the LDG method the boundary flux $\vec{\hat{q}}$ is defined as

$$
\vec{\hat{q}}_h = \{\vec{q}_h\} - \mathcal{E}[T_h] - \vec{\mathcal{C}}\,[\vec{q}_h]
$$

where $\mathcal{E}$ is a scalar (since $[T_h]$ is a vector) and $\vec{\mathcal{C}}$ is a vector (since $[\vec{q}_h]$ is a scalar). To be once again very explicit, the above equation writes as follows for a 2D Cartesian space:

$$
\begin{aligned}
\hat{q}_h{}_x &= \{q_h\}_x - \mathcal{E}[T_h]_x - \mathcal{C}_x[\vec{q}_h] \\
&= \frac{1}{2}(q_h{}_x + q_h{}_x) - \mathcal{E}(T_h n_x^+ + T_h n_x^-) - \mathcal{C}_x(\vec{q}_h \cdot \vec{n}^+ + \vec{q}_h \cdot \vec{n}^-) \tag{14.75} \\
\hat{q}_h{}_y &= \{q_h\}_y - \mathcal{E}[T_h]_y - \mathcal{C}_y[\vec{q}_h] \\
&= \frac{1}{2}(q_h{}_y + q_h{}_y) - \mathcal{E}(T_h n_y^+ + T_h n_y^-) - \mathcal{C}_y(\vec{q}_h \cdot \vec{n}^+ + \vec{q}_h \cdot \vec{n}^-) \tag{14.76}
\end{aligned}
$$

Remark. *In the book the note under table 4.1 states that the C_{ij} coefficients are constant matrices, which is quite misleading since some are actually scalars and others vectors.*

Filling Eqs. (14.75) and (14.76) into A of Eq. (14.73) leads to

$$
\begin{aligned}
A &= \int_{\partial\Omega} N_i^\theta \vec{q}_h \cdot \vec{n} \, dS \\
&= \int_{\partial\Omega} N_i^\theta \left[\{\vec{q}_h\} - \mathcal{E}[T_h] - \vec{\mathcal{C}}[\vec{q}_h] \right] \cdot \vec{n}^+ \, dS \\
&= \int_{\partial\Omega} N_i^\theta \left[\frac{1}{2}(\vec{q}_h + \vec{q}_h) - \mathcal{E}(T_h \vec{n}^+ + T_h \vec{n}^-) - \vec{\mathcal{C}}[\vec{q}_h] \right] \cdot \vec{n}^+ \, dS \\
&= \int_{\partial\Omega} N_i^\theta \left[\frac{1}{2}(\vec{q}_h + \vec{q}_h) \cdot \vec{n}^+ - \mathcal{E}(T_h \vec{n}^+ + T_h \vec{n}^-) \cdot \vec{n}^+ - \vec{\mathcal{C}} \cdot \vec{n}^+ [\vec{q}_h] \right] dS \\
&= \int_{\partial\Omega} N_i^\theta \left[\frac{1}{2}(\vec{q}_h + \vec{q}_h) \cdot \vec{n}^+ - \mathcal{E}(T_h - T_h) - (\vec{\mathcal{C}} \cdot \vec{n}^+)(\vec{q}_h - \vec{q}_h) \cdot \vec{n}^+ \right] dS \qquad \text{since} \quad \vec{n}^+ \cdot \vec{n}^+ = 1 \quad \vec{n}^+ \cdot \vec{n}^- = \\
&= \int_{\partial\Omega} N_i^\theta \left[\left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{q}_h \cdot \vec{n}^+ - \mathcal{E}T_h \right] dS + \int_{\partial\Omega} N_i^\theta \left[\left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{q}_h \cdot \vec{n}^+ + \mathcal{E}T_h \right] dS \\
&= \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{q}_h \cdot \vec{n}^+ \, dS - \int_{\partial\Omega} N_i^\theta \mathcal{E}T_h \, dS + \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{q}_h \cdot \vec{n}^+ \, dS + \int_{\partial\Omega} N_i^\theta \mathcal{E}T_h \, dS \\
&= A_1 + A_2 + A_3 + A_4 \tag{14.}
\end{aligned}
$$

In order to simplify notations we choose $N^q = N^\theta = N$ and drop the h subscripts.

$$A_1 = \int_{\partial\Omega} N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{q} \cdot \vec{n}^+ \, dS$$

$$= \int_{\partial\Omega} N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) (q_x n_x^+ + q_y n_y^+) \, dS$$

$$= \int_{\partial\Omega} N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) q_x n_x^+ \, dS + \int_{\partial\Omega} N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) q_y n_y^+ \, dS$$

$$\Rightarrow \vec{A}_1 = \left(\int_{\partial\Omega} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \, dS\right) \cdot \vec{q}_x + \left(\int_{\partial\Omega} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \, dS\right) \cdot \vec{q}_y \quad (14.78)$$

$$A_2 = -\int_{\partial\Omega} N_i \mathcal{E} T \, dS$$

$$\Rightarrow \vec{A}_2 = -\left(\int_{\partial\Omega} \mathcal{E} \vec{N}^T \vec{N} dS\right) \cdot \vec{T} \quad (14.79)$$

$$A_3 = \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{q} \cdot \vec{n}^+ \, dS$$

$$= \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) (q_x n_x^+ + q_y n_y^+) \, dS$$

$$= \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) q_x n_x^+ \, dS + \int_{\partial\Omega} N_i^\theta \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) q_y n_y^+ \, dS$$

$$\Rightarrow \vec{A}_3 = \left(\int_{\partial\Omega} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \, dS\right) \cdot \vec{q}_x + \left(\int_{\partial\Omega} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \, dS\right) \cdot \vec{q}_y \quad (14.80)$$

$$A_4 = \int_{\partial\Omega} N_i \mathcal{E} T \, dS$$

$$\Rightarrow \vec{A}_4 = \left(\int_{\partial\Omega} \mathcal{E} \vec{N}^T \vec{N} dS\right) \cdot \vec{T} \quad (14.81)$$

$$B = \int_\Omega \vec{\nabla} N_i \cdot \vec{q} \, dV$$

$$= \int_\Omega (\partial_x N_i q_x + \partial_y N_i q_y) \, dV$$

$$= \int_\Omega \partial_x N_i q_x \, dV + \int_\Omega \partial_y N_i q_y \, dV$$

$$\Rightarrow \vec{B} = \left(\int_\Omega \partial_x \vec{N}^T \vec{N} \, dV\right) \cdot \vec{q}_x + \left(\int_\Omega \partial_y \vec{N}^T \vec{N} \, dV\right) \cdot \vec{q}_y \quad (14.82)$$

$$C = \int_\Omega N_i H \, dV$$

$$\Rightarrow \vec{C} = \int_\Omega \vec{N}^T H \, dV \quad (14.83)$$

The expressions above find their equivalent in the book (NB stands for neighbour):

$$\left[\int_{\Omega_j} (\nabla\Phi)\Phi^T dV\right] \cdot \mathbf{q} \quad \left[\int_{\partial\Omega_j} (1/2 - C_{12})\Phi\Phi^T \mathbf{n} dS\right] \cdot \mathbf{q}$$

$$-\left[\int_{\partial\Omega_j} (1/2 + C_{12})\Phi\Phi^T \mathbf{n} dS\right] \cdot \mathbf{q}_{(NB)} \quad \left[\int_{\partial\Omega_j} C_{11}\Phi\Phi^T dS\right] \mathbf{T}$$

$$-\left[\int_{\partial\Omega_j} C_{11}\Phi\Phi^T dS\right] \mathbf{T}_{(NB)} \quad \int_{\Omega_j} \Phi Q dV \qquad (4.23b)$$

check minus signs

ODE 2 Its weak form writes:

$$\int_\Omega N_i^q (\vec{q} + k\vec{\nabla}T) dV = 0$$

or, (once again we drop the superscript on the basis functions and the h):

$$
\begin{aligned}
0 &= \int_\Omega N_i \vec{q} \, dV + \int_\Omega N_i k \vec{\nabla} T \, dV \\
&= \int_\Omega N_i \vec{q} \, dV + \int_\Omega \vec{\nabla}(N_i k T) \, dV - \int_\Omega \vec{\nabla}(k N_i) T \, dV
\end{aligned} \tag{14.84}
$$

We then use $\int_\Omega \vec{\nabla} f \, dV = \int_{\partial\Omega} f d\vec{S}$ and as before the temperature on the edge integral should be $\hat{T}$:

$$\int_\Omega N_i \vec{q} \, dV + \int_{\partial\Omega} k N_i \hat{T} \, \vec{n}^+ dS - \int_\Omega \vec{\nabla}(k N_i) T \, dV = 0$$

or, if decomposed in a 2D Cartesian axis system

$$
0 = \underbrace{\int_\Omega N_i q \, dV}_{D} + \underbrace{\int_{\partial\Omega} k N_i \hat{T} \, n_x^+ \, dS}_{E} - \underbrace{\int_\Omega \partial_x (k N_i) \, T \, dV}_{F} \qquad (0 = D + E - F)
$$

$$
0 = \underbrace{\int_\Omega N_i q \, dV}_{G} + \underbrace{\int_{\partial\Omega} k N_i \hat{T} \, n_y^+ \, dS}_{H} - \underbrace{\int_\Omega \partial_y (k N_i) \, T \, dV}_{I} \qquad (0 = G + H - I) \tag{14.85}
$$

which (aside from a minus sign coming from a different definition of the heat flux) is 'identical' to the book (although the notations in the book are hella confusing):

$$\int_{\Omega_j} \mathbf{q}_h \cdot \mathbf{w} \, dV = -\int_{\Omega_j} T_h \nabla \cdot (\kappa \mathbf{w}) dV + \int_{\partial\Omega_j} \kappa \hat{T}_h \mathbf{n}_j \cdot \mathbf{w} dS$$
$$\text{(with } \kappa \to k, \, \boldsymbol{w} \to \vec{N})$$

The temperature flux $\hat{T}$ is chosen to be (table 4.1 in book):

$$
\begin{aligned}
\hat{T}_h &= \{T_h\} + \vec{\mathcal{C}}[T_h] - \mathcal{F}[\vec{q}_h] \tag{14.86} \\
&= \frac{1}{2}(T_h + T_h) + \vec{\mathcal{C}}(T_h \vec{n}^+ + T_h \vec{n}^-) - \mathcal{F}(\vec{q}_h \cdot \vec{n}^+ + \vec{q}_h \cdot \vec{n}^-) \tag{14.87} \\
&= \frac{1}{2}(T_h + T_h) + (T_h - T_h) \, \vec{\mathcal{C}} \cdot \vec{n}^+ - \mathcal{F}(\vec{q}_h - \vec{q}_h) \cdot \vec{n}^+ \tag{14.88} \\
&= \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) T_h + \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) T_h - \mathcal{F}(\vec{q}_h - \vec{q}_h) \cdot \vec{n}^+ \tag{14.89}
\end{aligned}
$$

In what follows we assume k to be constant within each element so that $\partial_x(k N_i) = k \partial_x N_i$.

$$D = \int_\Omega N_i \underset{x}{q}\, dV \qquad \Rightarrow \qquad \vec{D} = \left(\int_\Omega \vec{N}^T \vec{N} dV \right) \cdot \underset{x}{\vec{q}} \tag{14.90}$$

$$F = \int_\Omega k \partial_x N_i\, T\, dV \qquad \Rightarrow \qquad \vec{F} = \left(\int_\Omega k \partial_x \vec{N}^T \vec{N}\, dV \right) \cdot \vec{T} \tag{14.91}$$

$$G = \int_\Omega N_i \underset{y}{q^h}\, dV \qquad \Rightarrow \qquad \vec{G} = \left(\int_\Omega \vec{N}^T \vec{N} dV \right) \cdot \underset{y}{\vec{q}} \tag{14.92}$$

$$I = \int_\Omega k \partial_y N_i\, T\, dV \qquad \Rightarrow \qquad \vec{I} = \left(\int_\Omega k \partial_y \vec{N}^T \vec{N}\, dV \right) \cdot \vec{T} \tag{14.93}$$

$$
\begin{aligned}
E &= \int_{\partial\Omega} k N_i \hat{T}\, n_x^+ dS \\
&= \int_{\partial\Omega} k N_i \left[\left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h + \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h - \mathcal{F}(\vec{q}_h - \vec{q}_h) \cdot \vec{n}^+ \right] n_x^+ dS \\
&= \int_{\partial\Omega} k N_i \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h n_x^+ dS + \int_{\partial\Omega} k N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h n_x^+ dS \\
&\quad - \int_{\partial\Omega} k N_i \mathcal{F}(\vec{q}_h \cdot \vec{n}^+) n_x^+ dS + \int_{\partial\Omega} k N_i \mathcal{F}(\vec{q}_h \cdot \vec{n}^+) n_x^+ dS \\
&= E_1 + E_2 + E_3 + E_4 \\
H &= \int_{\partial\Omega} k N_i \hat{T}\, n_y^+ dS \\
&= H_1 + H_2 + H_3 + H_4
\end{aligned} \tag{14.94}
$$

Then

$$E_1 = \int_{\partial\Omega} k N_i \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h n_x^+ dS \qquad \Rightarrow \qquad \left(\int_{\partial\Omega} k \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_x^+ dS \right) \cdot \vec{T}$$

$$E_2 = \int_{\partial\Omega} k N_i \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) T_h n_x^+ dS \qquad \Rightarrow \qquad \left(\int_{\partial\Omega} k \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_x^+ dS \right) \cdot \vec{T}$$

$$
\begin{aligned}
E_3 &= - \int_{\partial\Omega} k N_i \mathcal{F}(\vec{q}_h \cdot \vec{n}^+) n_x^+ dS \\
&= - \int_{\partial\Omega} k N_i \mathcal{F}(q_x n_x^+ + q_y n_y^+) n_x^+ dS \\
&= - \int_{\partial\Omega} k N_i \mathcal{F} q_x n_x^+ n_x^+ dS - \int_{\partial\Omega} k N_i \mathcal{F} q_y n_y^+ n_x^+ dS \\
&\Rightarrow - \left(\int_{\partial\Omega} k \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS \right) \cdot \underset{x}{\vec{q}} - \left(\int_{\partial\Omega} k \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS \right) \cdot \underset{y}{\vec{q}} \\
E_4 &= \int_{\partial\Omega} k N_i \mathcal{F}(\vec{q}_h \cdot \vec{n}^+) n_x^+ dS \\
&= \int_{\partial\Omega} k N_i \mathcal{F}(q_x n_x^+ + q_y n_y^+) n_x^+ dS \\
&= \int_{\partial\Omega} k N_i \mathcal{F} q_x n_x^+ n_x^+ dS + \int_{\partial\Omega} k N_i \mathcal{F} q_y n_y^+ n_x^+ dS \\
&\Rightarrow \left(\int_{\partial\Omega} k \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS \right) \cdot \underset{x}{\vec{q}} + \left(\int_{\partial\Omega} k \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS \right) \cdot \underset{y}{\vec{q}}
\end{aligned} \tag{14.95}
$$

The H_i terms are so similar to the E_i terms that there is not need to write them out explicitly.

We have seen that ODE #2 and #1 write

$$D + E - F = D + E_1 + E_2 + E_3 + E_4 - F = 0 \tag{14.96}$$

$$G + H - I = G + H_1 + H_2 + H_3 + H_4 - I = 0 \tag{14.97}$$

$$A - B = A_1 + A_2 + A_3 + A_4 - B = \vec{C} \tag{14.98}$$

so that

$$\left(\int_\Omega \vec{N}^T \vec{N} dV\right) \cdot \vec{q}_x + \left(\int_{\partial\Omega} k\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_x^+ dS\right)\cdot \vec{T} + \left(\int_{\partial\Omega} k\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_x^+ dS\right)\cdot \vec{T}$$

$$- \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_x^+ dS\right)\cdot \vec{q}_x - \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_x^+ dS\right)\cdot \vec{q}_x + \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_x^+ dS\right)\cdot \vec{q}_y + \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_x^+ dS\right)\cdot \vec{q}_y - \left(\int_\Omega k\partial_x \vec{N}^T \vec{N}\, dV\right)\cdot \vec{T} =$$

$$\left(\int_\Omega \vec{N}^T \vec{N} dV\right)\cdot \vec{q}_y + \left(\int_{\partial\Omega} k\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_y^+ dS\right)\cdot \vec{T} + \left(\int_{\partial\Omega} k\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_y^+ dS\right)\cdot \vec{T}$$

$$- \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_y^+ dS\right)\cdot \vec{q}_x - \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_y^+ dS\right)\cdot \vec{q}_x + \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_y^+ dS\right)\cdot \vec{q}_y + \left(\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_y^+ dS\right)\cdot \vec{q}_y - \left(\int_\Omega k\partial_y \vec{N}^T \vec{N}\, dV\right)\cdot \vec{T} =$$

$$\left(\int_{\partial\Omega}\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_x^+\, dS\right)\cdot \vec{q}_x + \left(\int_{\partial\Omega}\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_y^+\, dS\right)\cdot \vec{q}_y - \left(\int_{\partial\Omega}\mathcal{E}\vec{N}^T \vec{N} dS\right)\cdot \vec{T}$$

$$+ \left(\int_{\partial\Omega}\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_x^+\, dS\right)\cdot \vec{q}_x + \left(\int_{\partial\Omega}\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_y^+\, dS\right)\cdot \vec{q}_y + \left(\int_{\partial\Omega}\mathcal{E}\vec{N}^T \vec{N} dS\right)\cdot \vec{T} - \left(\int_\Omega \partial_x \vec{N}^T \vec{N}\, dV\right)\cdot \vec{q}_x - \left(\int_\Omega \partial_y \vec{N}^T \vec{N}\, dV\right)\cdot \vec{q}_y =$$

$$\left[\underbrace{\begin{pmatrix} \int_\Omega \vec{N}^T \vec{N} dV & 0 & -\int_\Omega k\partial_x \vec{N}^T \vec{N}\, dV \\ 0 & \int_\Omega \vec{N}^T \vec{N} dV & -\int_\Omega k\partial_y \vec{N}^T \vec{N}\, dV \\ -\int_\Omega \partial_x \vec{N}^T \vec{N}\, dV & -\int_\Omega \partial_y \vec{N}^T \vec{N}\, dV & 0 \end{pmatrix}}_{\mathcal{A}_\Omega} + \underbrace{\begin{pmatrix} -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_x^+ dS & -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_x^+ dS & \int_{\partial\Omega} k\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_x^+ dS \\ -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_y^+ dS & -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_y^+ dS & \int_{\partial\Omega} k\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_y^+ dS \\ \int_{\partial\Omega}\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_x^+\, dS & \int_{\partial\Omega}\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_y^+\, dS & -\int_{\partial\Omega}\mathcal{E}\vec{N}^T \vec{N} dS \end{pmatrix}}_{\mathcal{A}_{\partial\Omega}}\right]\cdot \begin{pmatrix} \vec{q}_x \\ \vec{q}_y \\ \vec{T} \end{pmatrix}$$

$$= \begin{pmatrix} -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_x^+ dS & -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_x^+ dS & -\int_{\partial\Omega} k\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_x^+ dS \\ -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_x^+ n_y^+ dS & -\int_{\partial\Omega} k\vec{N}^T \vec{N}\mathcal{F} n_y^+ n_y^+ dS & -\int_{\partial\Omega} k\left(\frac{1}{2} - \vec{C}\cdot\vec{n}^+\right) N^T \vec{N} n_y^+ dS \\ -\int_{\partial\Omega}\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_x^+\, dS & -\int_{\partial\Omega}\left(\frac{1}{2} + \vec{C}\cdot\vec{n}^+\right) \vec{N}^T \vec{N} n_y^+\, dS & -\int_{\partial\Omega}\mathcal{E}\vec{N}^T \vec{N} dS \end{pmatrix}\cdot \begin{pmatrix} \vec{q}_x \\ \vec{q}_y \\ \vec{T} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \int_\Omega \vec{N}^T H\, dV \end{pmatrix}$$

$$
\left[\underbrace{\begin{pmatrix} \int_\Omega \vec{N}^T \vec{N} dV & 0 & -\int_\Omega k\partial_x \vec{N}^T \vec{N}\ dV \\ 0 & \int_\Omega \vec{N}^T \vec{N} dV & -\int_\Omega k\partial_y \vec{N}^T \vec{N}\ dV \\ -\int_\Omega \partial_x \vec{N}^T \vec{N}\ dV & -\int_\Omega \partial_y \vec{N}^T \vec{N}\ dV & 0 \end{pmatrix}}_{\mathcal{A}_\Omega} + \sum_{i=1}^{nedges} \underbrace{\begin{pmatrix} -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_x^+ n_x^+ dS & -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_y^+ n_x^+ dS & \int_{\partial\Omega_i} k\left(\frac{1}{2}+\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_x^+ dS \\ -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_x^+ n_y^+ dS & -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_y^+ n_y^+ dS & \int_{\partial\Omega_i} k\left(\frac{1}{2}+\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_y^+ dS \\ \int_{\partial\Omega_i} \left(\frac{1}{2}-\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_x^+\ dS & \int_{\partial\Omega_i} \left(\frac{1}{2}-\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_y^+\ dS & -\int_{\partial\Omega_i} \mathcal{E}\vec{N}^T \vec{N}dS \end{pmatrix}}_{\mathcal{A}_{\partial\Omega}} \right] \cdot
$$

$$
= \sum_{i=1}^{nedges} \begin{pmatrix} -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_x^+ n_x^+ dS & -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_y^+ n_x^+ dS & -\int_{\partial\Omega_i} k\left(\frac{1}{2}-\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_x^+ dS \\ -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_x^+ n_y^+ dS & -\int_{\partial\Omega_i} k\vec{N}^T \vec{N}\mathcal{F}n_y^+ n_y^+ dS & -\int_{\partial\Omega_i} k\left(\frac{1}{2}-\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_y^+ dS \\ -\int_{\partial\Omega_i} \left(\frac{1}{2}+\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_x^+\ dS & -\int_{\partial\Omega_i} \left(\frac{1}{2}+\vec{\mathcal{C}}\cdot\vec{n}^+\right) N^T \vec{N}n_y^+\ dS & -\int_{\partial\Omega_i} \mathcal{E}\vec{N}^T \vec{N}dS \end{pmatrix} \cdot \begin{pmatrix} \vec{q}_x \\ \vec{q}_y \\ \vec{T} \end{pmatrix}_i + \begin{pmatrix} 0 \\ 0 \\ \int_\Omega \vec{N}^T H\ dV \end{pmatrix}
$$

$$
\left[\begin{pmatrix} E & 0 & H_x \\ 0 & E & H_y \\ J_x & J_y & 0 \end{pmatrix} + \sum_{i=1}^{nedges} \begin{pmatrix} E_{xx,i} & E_{xy,i} & H_{x,i} \\ E_{yx,i} & E_{yy,i} & H_{y,i} \\ J_{x,i} & J_{y,i} & G_{T,i} \end{pmatrix} \right] \cdot \begin{pmatrix} \vec{q}_x \\ \vec{q}_y \\ \vec{T} \end{pmatrix} =
$$

which is identical to the equation 4.24 in Li's book (if the terms related to the third dimension are disregarded):

$$
\begin{bmatrix} E & 0 & 0 & H_x \\ 0 & E & 0 & H_y \\ 0 & 0 & E & H_z \\ J_x & J_y & J_z & 0 \end{bmatrix} \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_y \\ \mathbf{q}_z \\ \underline{T} \end{bmatrix} + \sum_{i=1}^{NS} \begin{bmatrix} E_{xx,i} & E_{xy,i} & E_{xz,i} & H_{x,i} \\ E_{yx,i} & E_{yy,i} & E_{yz,i} & H_{y,i} \\ E_{zx,i} & E_{zy,i} & E_{zz,i} & H_{z,i} \\ J_{x,i} & J_{y,i} & J_{z,i} & G_{T,i} \end{bmatrix} \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_y \\ \mathbf{q}_z \\ \underline{T} \end{bmatrix}
$$
$$
+ \sum_{i=1}^{NS} \begin{bmatrix} E_{xx,B,i} & E_{xy,B,i} & E_{xy,B,i} & H_{x,B,i} \\ E_{yx,B,i} & E_{yy,B,i} & E_{yz,B,i} & H_{y,B,i} \\ E_{zx,B,i} & E_{zy,B,i} & E_{zz,B,i} & H_{z,B,i} \\ J_{x,B,i} & J_{y,B,i} & J_{z,B,i} & G_{T,B,i} \end{bmatrix} \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_y \\ \mathbf{q}_z \\ \underline{T} \end{bmatrix}_{(NB,i)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{S}_T \end{bmatrix}
$$

$$\boldsymbol{E} \quad = \quad \int_{\Omega} \vec{N}^T \vec{N} dV \tag{14.99}$$

$$\boldsymbol{H}_x \quad = \quad -\int_{\Omega} k\partial_x \vec{N}^T \vec{N} \, dV \tag{14.100}$$

$$\boldsymbol{H}_y \quad = \quad -\int_{\Omega} k\partial_y \vec{N}^T \vec{N} \, dV \tag{14.101}$$

$$\boldsymbol{J}_x \quad = \quad -\int_{\Omega} \partial_x \vec{N}^T \vec{N} \, dV \tag{14.102}$$

$$\boldsymbol{J}_y \quad = \quad -\int_{\Omega} \partial_y \vec{N}^T \vec{N} \, dV \tag{14.103}$$

$$\boldsymbol{E}_{xx,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS \tag{14.104}$$

$$\boldsymbol{E}_{xy,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS \tag{14.105}$$

$$\boldsymbol{E}_{yx,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_x^+ n_y^+ dS \tag{14.106}$$

$$\boldsymbol{E}_{yy,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_y^+ n_y^+ dS \tag{14.107}$$

$$\boldsymbol{H}_{x,i} \quad = \quad \int_{\partial\Omega_i} k \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_x^+ dS \tag{14.108}$$

$$\boldsymbol{H}_{y,i} \quad = \quad \int_{\partial\Omega_i} k \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_y^+ dS \tag{14.109}$$

$$\boldsymbol{J}_{x,i} \quad = \quad \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{N}^T \vec{N} n_x^+ \, dS \tag{14.110}$$

$$\boldsymbol{J}_{y,i} \quad = \quad \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{N}^T \vec{N} n_y^+ \, dS \tag{14.111}$$

$$\boldsymbol{G}_{T,i} \quad = \quad -\int_{\partial\Omega_i} \mathcal{E} \vec{N}^T \vec{N} dS \tag{14.112}$$

$$\boldsymbol{E}_{xx,B,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS \tag{14.113}$$

$$\boldsymbol{E}_{xy,B,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS \tag{14.114}$$

$$\boldsymbol{E}_{yx,B,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_x^+ n_y^+ dS \tag{14.115}$$

$$\boldsymbol{E}_{yy,B,i} \quad = \quad -\int_{\partial\Omega_i} k\vec{N}^T \vec{N} \mathcal{F} n_y^+ n_y^+ dS \tag{14.116}$$

$$\boldsymbol{H}_{x,B,i} \quad = \quad -\int_{\partial\Omega_i} k \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_x^+ dS \tag{14.117}$$

$$\boldsymbol{H}_{y,B,i} \quad = \quad -\int_{\partial\Omega_i} k \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+ \right) N^T \vec{N} n_y^+ dS \tag{14.118}$$

$$\boldsymbol{J}_{x,B,i} \quad = \quad -\int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{N}^T \vec{N} n_x^+ \, dS \tag{14.119}$$

$$\boldsymbol{J}_{y,B,i} \quad = \quad -\int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+ \right) \vec{N}^T \vec{N} n_y^+ \, dS \tag{14.120}$$

$$\boldsymbol{G}_{T,B,i} \quad = \quad -\int_{\partial\Omega_i} \mathcal{E} \vec{N}^T \vec{N} dS \tag{14.121}$$

$$\boldsymbol{S}_T \quad = \quad \int_{\Omega} \vec{N}^T H \, dV \tag{14.122}$$

Note that $\mathcal{E} = C_{11}$ and $\mathcal{F} = C_{22}$ in the book.

If k is constant per element, then:

$$\boldsymbol{E} = \int_\Omega \vec{N}^T \vec{N} dV \tag{14.123}$$

$$\boldsymbol{H}_x = -k \int_\Omega \partial_x \vec{N}^T \vec{N} \; dV \tag{14.124}$$

$$\boldsymbol{H}_y = -k \int_\Omega \partial_y \vec{N}^T \vec{N} \; dV \tag{14.125}$$

$$\boldsymbol{J}_x = -\int_\Omega \partial_x \vec{N}^T \vec{N} \; dV \tag{14.126}$$

$$\boldsymbol{J}_y = -\int_\Omega \partial_y \vec{N}^T \vec{N} \; dV \tag{14.127}$$

$$\boldsymbol{E}_{xx,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS \tag{14.128}$$

$$\boldsymbol{E}_{xy,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS \tag{14.129}$$

$$\boldsymbol{E}_{yx,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_y^+ dS \tag{14.130}$$

$$\boldsymbol{E}_{yy,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_y^+ dS \tag{14.131}$$

$$\boldsymbol{H}_{x,i} = k \int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) N^T \vec{N} n_x^+ dS = -k\boldsymbol{J}_{x,B,i} \tag{14.132}$$

$$\boldsymbol{H}_{y,i} = k \int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) N^T \vec{N} n_y^+ dS = -k\boldsymbol{J}_{y,B,i} \tag{14.133}$$

$$\boldsymbol{J}_{x,i} = \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS \tag{14.134}$$

$$\boldsymbol{J}_{y,i} = \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \; dS \tag{14.135}$$

$$\boldsymbol{G}_{T,i} = -\int_{\partial\Omega_i} \mathcal{E} \vec{N}^T \vec{N} dS \tag{14.136}$$

$$\boldsymbol{E}_{xx,B,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_x^+ dS = \boldsymbol{E}_{xx,i} \tag{14.137}$$

$$\boldsymbol{E}_{xy,B,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_x^+ dS = \boldsymbol{E}_{xy,i} \tag{14.138}$$

$$\boldsymbol{E}_{yx,B,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_x^+ n_y^+ dS = \boldsymbol{E}_{yx,i} \tag{14.139}$$

$$\boldsymbol{E}_{yy,B,i} = -k \int_{\partial\Omega_i} \vec{N}^T \vec{N} \mathcal{F} n_y^+ n_y^+ dS = \boldsymbol{E}_{yy,i} \tag{14.140}$$

$$\boldsymbol{H}_{x,B,i} = -k \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) N^T \vec{N} n_x^+ dS = -k\boldsymbol{J}_{x,i} \tag{14.141}$$

$$\boldsymbol{H}_{y,B,i} = -k \int_{\partial\Omega_i} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) N^T \vec{N} n_y^+ dS = -k\boldsymbol{J}_{y,i} \tag{14.142}$$

$$\boldsymbol{J}_{x,B,i} = -\int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS \tag{14.143}$$

$$\boldsymbol{J}_{y,B,i} = -\int_{\partial\Omega_i} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \; dS \tag{14.144}$$

$$\boldsymbol{G}_{T,B,i} = -\int_{\partial\Omega_i} \mathcal{E} \vec{N}^T \vec{N} dS = \boldsymbol{G}_{T,i} \tag{14.145}$$

$$\boldsymbol{S}_T = \int_\Omega \vec{N}^T H \; dV \tag{14.146}$$

14.5.1 The special case of linear rectangular elements

Let us start with $\mathcal{A}_e$ where we assume that k is constant within an element:

$$\mathcal{A}_{\Omega_e} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{0} & \boldsymbol{H}_x \\ \boldsymbol{0} & \boldsymbol{E} & \boldsymbol{H}_y \\ \boldsymbol{J}_x & \boldsymbol{J}_y & \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{0} & -k_e\boldsymbol{J}_x \\ \boldsymbol{0} & \boldsymbol{E} & -k_e\boldsymbol{J}_y \\ \boldsymbol{J}_x & \boldsymbol{J}_y & \boldsymbol{0} \end{pmatrix} \qquad (14.147)$$

The matrices $\boldsymbol{E}$, $\boldsymbol{H}_x$, and $\boldsymbol{H}_y$ have been analytically derived in Appendix C.0.2:

$$\boldsymbol{E} = \frac{h_x h_y}{9} \begin{pmatrix} 1 & 1/2 & 1/4 & 1/2 \\ 1/2 & 1 & 1/2 & 1/4 \\ 1/4 & 1/2 & 1 & 1/2 \\ 1/2 & 1/4 & 1/2 & 1 \end{pmatrix} \quad \boldsymbol{J}_x = \frac{h_y}{12} \begin{pmatrix} -2 & -2 & -1 & -1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ -1 & -1 & -2 & -2 \end{pmatrix} \quad \boldsymbol{J}_y = \frac{h_x}{12} \begin{pmatrix} -2 & -1 & -1 & -2 \\ -1 & -2 & -2 & -1 \\ 1 & 2 & 2 & 1 \\ 2 & 1 & 1 & 2 \end{pmatrix}$$

The matrix $\mathcal{A}_{\Omega_e}$ is therefore trivial to implement.

Let us now turn to $\mathcal{A}_{\partial\Omega}$ which is specific to the DG method. Because elements are rectangles then $n_i^+ n_j^+ = 0$ if $i \neq j$ (where i and j are edge indices). Also, if $i = j$ then $n_i^+ n_j^+ = 1$.
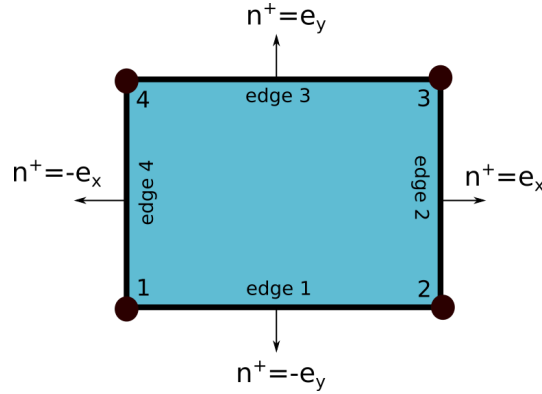
Assuming here again that heat conductivities are constant inside an element it then follows that

$$\mathcal{A}_{\partial\Omega_e} = \sum_{i=1}^{nedges} \begin{pmatrix} \boldsymbol{E}_{xx,i} & \boldsymbol{E}_{xy,i} & \boldsymbol{H}_{x,i} \\ \boldsymbol{E}_{yx,i} & \boldsymbol{E}_{yy,i} & \boldsymbol{H}_{y,i} \\ \boldsymbol{J}_{x,i} & \boldsymbol{J}_{y,i} & \boldsymbol{G}_{T,i} \end{pmatrix} = \sum_{i=1}^{nedges} \begin{pmatrix} \boldsymbol{E}_{xx,i} & \boldsymbol{0} & \boldsymbol{H}_{x,i} \\ \boldsymbol{0} & \boldsymbol{E}_{yy,i} & \boldsymbol{H}_{y,i} \\ \boldsymbol{J}_{x,i} & \boldsymbol{J}_{y,i} & \boldsymbol{G}_{T,i} \end{pmatrix} \qquad (14.148)$$

Note that

- $i = 1$: bottom edge, i.e. $s = -1$ and then $N_4 = N_3 = 0$; Also $\vec{n}_x^+ = 0$, $\vec{n}_y^+ = -1$

- $i = 2$: right edge, i.e. $r = +1$ and then $N_1 = N_4 = 0$; Also $\vec{n}_x^+ = 1$, $\vec{n}_y^+ = 0$

- $i = 3$: top edge, i.e. $s = +1$ and then $N_1 = N_2 = 0$; Also $\vec{n}_x^+ = 0$, $\vec{n}_y^+ = 1$

- $i = 4$: left edge, i.e. $r = -1$ and then $N_2 = N_3 = 0$; Also $\vec{n}_x^+ = -1$, $\vec{n}_y^+ = 0$

Then

$$\boldsymbol{G}_{T,1} = -\mathcal{E} \int_{\partial\Omega_1} \vec{N}^T \vec{N} dS = -\mathcal{E}\boldsymbol{C}_1$$

$$\boldsymbol{G}_{T,2} = -\mathcal{E} \int_{\partial\Omega_2} \vec{N}^T \vec{N} dS = -\mathcal{E}\boldsymbol{C}_2$$

$$\boldsymbol{G}_{T,3} = -\mathcal{E} \int_{\partial\Omega_3} \vec{N}^T \vec{N} dS = -\mathcal{E}\boldsymbol{C}_3$$

$$\boldsymbol{G}_{T,4} = -\mathcal{E} \int_{\partial\Omega_4} \vec{N}^T \vec{N} dS = -\mathcal{E}\boldsymbol{C}_4$$

where the matrices $\boldsymbol{C}_i$ have been worked out in detail in appendix C.0.2:

$$\boldsymbol{C}_1 = \int_{1\to 2} \vec{N}^T \vec{N} dS = \frac{h_x}{6} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{C}_2 = \int_{2\to 3} \vec{N}^T \vec{N} dS = \frac{h_y}{6} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{C}_3 = \int_{3\to 4} \vec{N}^T \vec{N} dS = \frac{h_x}{6} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

$$\boldsymbol{C}_4 = \int_{4\to 1} \vec{N}^T \vec{N} dS = \frac{h_y}{6} \begin{pmatrix} 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}$$

$$\boldsymbol{E}_{xx,1} = -k_e \mathcal{F} \int_{\partial\Omega_1} \vec{N}^T \vec{N} n_x^+ n_x^+ dS = 0$$

$$\boldsymbol{E}_{xx,2} = -k_e \mathcal{F} \int_{\partial\Omega_2} \vec{N}^T \vec{N} n_x^+ n_x^+ dS = -k_e \mathcal{F} \boldsymbol{C}_2$$

$$\boldsymbol{E}_{xx,3} = -k_e \mathcal{F} \int_{\partial\Omega_3} \vec{N}^T \vec{N} n_x^+ n_x^+ dS = 0$$

$$\boldsymbol{E}_{xx,4} = -k_e \mathcal{F} \int_{\partial\Omega_4} \vec{N}^T \vec{N} n_x^+ n_x^+ dS = -k_e \mathcal{F} \boldsymbol{C}_4$$

$$\boldsymbol{E}_{yy,1} = -k_e \mathcal{F} \int_{\partial\Omega_1} \vec{N}^T \vec{N} n_y^+ n_y^+ dS = -k_e \mathcal{F} \boldsymbol{C}_1$$

$$\boldsymbol{E}_{yy,2} = -k_e \mathcal{F} \int_{\partial\Omega_2} \vec{N}^T \vec{N} n_y^+ n_y^+ dS = 0$$

$$\boldsymbol{E}_{yy,3} = -k_e \mathcal{F} \int_{\partial\Omega_3} \vec{N}^T \vec{N} n_y^+ n_y^+ dS = -k_e \mathcal{F} \boldsymbol{C}_3$$

$$\boldsymbol{E}_{yy,4} = -k_e \mathcal{F} \int_{\partial\Omega_4} \vec{N}^T \vec{N} n_y^+ n_y^+ dS = 0$$

$$\boldsymbol{H}_{x,1} = k_e \int_{\partial\Omega_1} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ dS = 0$$

$$\boldsymbol{H}_{x,2} = k_e \int_{\partial\Omega_2} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ dS = k_e \left(\frac{1}{2} + \mathcal{C}_x\right) \boldsymbol{C}_2$$

$$\boldsymbol{H}_{x,3} = k_e \int_{\partial\Omega_3} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ dS = 0$$

$$\boldsymbol{H}_{x,4} = k_e \int_{\partial\Omega_4} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ dS = -k_e \left(\frac{1}{2} - \mathcal{C}_x\right) \boldsymbol{C}_4$$

$$\boldsymbol{H}_{y,1} = k_e \int_{\partial\Omega_1} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ dS = -k_e \left(\frac{1}{2} - \mathcal{C}_y\right) \boldsymbol{C}_1$$

$$\boldsymbol{H}_{y,2} = k_e \int_{\partial\Omega_2} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ dS = 0$$

$$\boldsymbol{H}_{y,3} = k_e \int_{\partial\Omega_3} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ dS = k_e \left(\frac{1}{2} + \mathcal{C}_y\right) \boldsymbol{C}_3$$

$$\boldsymbol{H}_{y,4} = k_e \int_{\partial\Omega_4} \left(\frac{1}{2} + \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ dS = 0$$

$$\boldsymbol{J}_{x,1} = \int_{\partial\Omega_1} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS = 0$$

$$\boldsymbol{J}_{x,2} = \int_{\partial\Omega_2} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS = \left(\frac{1}{2} - \mathcal{C}_x\right) \boldsymbol{C}_2$$

$$\boldsymbol{J}_{x,3} = \int_{\partial\Omega_3} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS = 0$$

$$\boldsymbol{J}_{x,4} = \int_{\partial\Omega_4} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_x^+ \; dS = -\left(\frac{1}{2} + \mathcal{C}_x\right) \boldsymbol{C}_4$$

$$\boldsymbol{J}_{y,1} = \int_{\partial\Omega_1} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \; dS = -\left(\frac{1}{2} + \mathcal{C}_y\right) \boldsymbol{C}_1$$

$$\boldsymbol{J}_{y,2} = \int_{\partial\Omega_2} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \; dS = 0$$

$$\boldsymbol{J}_{y,3} = \int_{\partial\Omega_3} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \overset{611}{\vec{N}^T} \vec{N} n_y^+ \; dS = \left(\frac{1}{2} - \mathcal{C}_y\right) \boldsymbol{C}_3$$

$$\boldsymbol{J}_{y,4} = \int_{\partial\Omega_4} \left(\frac{1}{2} - \vec{\mathcal{C}} \cdot \vec{n}^+\right) \vec{N}^T \vec{N} n_y^+ \; dS = 0$$

14.5.2 The special case of linear triangular elements
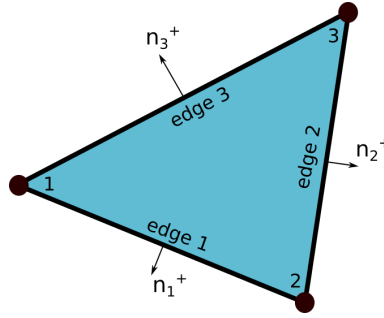
The linear basis functions in the triangle are

$$
\begin{aligned}
N_1(x,y) &= \frac{1}{2S}(x_2 y_3 - x_3 y_2 + (y_2 - y_3)x + (x_3 - x_2)y)\\
N_2(x,y) &= \frac{1}{2S}(x_3 y_1 - x_1 y_3 + (y_3 - y_1)x + (x_1 - x_3)y)\\
N_3(x,y) &= \frac{1}{2S}(x_1 y_2 - x_2 y_1 + (y_1 - y_2)x + (x_2 - x_1)y)
\end{aligned}
$$

where S is the area of the element:

$$
S = \frac{1}{2}[(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)]
$$

We can easily verifiy that $N_i(x_j, y_j) = \delta_{ij}$. We then have

$$
\begin{aligned}
\partial_x N_1(x,y) &= \frac{1}{2S}(y_2 - y_3)\\
\partial_x N_2(x,y) &= \frac{1}{2S}(y_3 - y_1)\\
\partial_x N_3(x,y) &= \frac{1}{2S}(y_1 - y_2)
\end{aligned}
$$

and

$$
\begin{aligned}
\partial_y N_1(x,y) &= \frac{1}{2S}(x_3 - x_2)\\
\partial_y N_2(x,y) &= \frac{1}{2S}(x_1 - x_3)\\
\partial_y N_3(x,y) &= \frac{1}{2S}(x_2 - x_1)
\end{aligned}
$$

Then, as shown in Section C.0.5, the mass matrix[3] and the $\boldsymbol{J}_x$ and $\boldsymbol{J}_y$ matrices are:

$$
\boldsymbol{E} = \int_\Omega \vec{N}^T \vec{N} dV = \frac{S}{12}\begin{pmatrix} 2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2 \end{pmatrix} \tag{14.149}
$$

$$
\boldsymbol{J}_x = -\int_\Omega \partial_x \vec{N}^T \vec{N}\, dV = -\frac{1}{6}\begin{pmatrix} y_2 - y_3 & y_2 - y_3 & y_2 - y_3\\ y_3 - y_1 & y_3 - y_1 & y_3 - y_1\\ y_1 - y_2 & y_1 - y_2 & y_1 - y_2 \end{pmatrix}
$$

$$
\boldsymbol{J}_y = -\int_\Omega \partial_y \vec{N}^T \vec{N}\, dV = -\frac{1}{6}\begin{pmatrix} x_3 - x_2 & x_3 - x_2 & x_3 - x_2\\ x_1 - x_3 & x_1 - x_3 & x_1 - x_3\\ x_2 - x_1 & x_2 - x_1 & x_2 - x_1 \end{pmatrix} \tag{14.150}
$$

---

[3]The mass matrix is commonly called $\boldsymbol{M}$ but I use here the same notations as in Li's book.

In the same appendix we show that

$$
\boldsymbol{C}_1 = \int_{\partial\Omega_3} \vec{N}^T \vec{N} dS = \frac{L_1}{6} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}
\tag{14.151}
$$

$$
\boldsymbol{C}_2 = \int_{\partial\Omega_1} \vec{N}^T \vec{N} dS = \frac{L_2}{6} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}
\tag{14.152}
$$

$$
\boldsymbol{C}_3 = \int_{\partial\Omega_2} \vec{N}^T \vec{N} dS = \frac{L_3}{6} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix}
\tag{14.153}
$$

Testing the waters - constant temperature field

Let us assume that the temperature is constant in space. It then follows that the heat flux is identically zero.

Testing the waters - linear temperature field

If the temperature field is given by $T(x,y) = T_0 - ax - by$ then $q_x = a$ and $q_y = b$.

14.6 Time-dependent diffusion PDE in 2D

14.7 Stokes equations

Two relevant papers:

- Cockburn *et al.* (2002) [283] - LDG

- Cockburn *et al.* (2010) [286] - HDG

Let us start with the dimensionless Stokes system [283]:

$$\begin{aligned}
-\eta\Delta\vec{\mathbf{v}} + \vec{\nabla}p &= \vec{f} &&\text{in } \Omega &&(14.154)\\
\vec{\nabla}\cdot\vec{\mathbf{v}} &= 0 &&\text{in } \Omega &&(14.155)\\
\vec{\mathbf{v}} &= \vec{\mathbf{v}}_D &&\text{on } \Gamma &&(14.156)
\end{aligned}$$

where Ω is a bounded domain of $\mathbb{R}^d$ and the Dirichlet boundary conditions are such that they satisfy the compatibility condition

$$\int_\Gamma \vec{\mathbf{v}}_D \cdot \vec{n} = 0$$

where $\vec{n}$ is the outward unit normal.

Gradient-based formulation In order to obtain the LDG methods we first rewrite this system as the following collection of conservation laws [283]:

$$\begin{aligned}
\boldsymbol{L} &= \vec{\nabla}\vec{\mathbf{v}} &&\text{in } \Omega &&(14.157)\\
\vec{\nabla}\cdot(-2\eta\boldsymbol{L} + p\mathbf{1}) &= \vec{f} &&\text{in } \Omega &&(14.158)\\
\vec{\nabla}\cdot\vec{\mathbf{v}} &= 0 &&\text{in } \Omega &&(14.159)\\
\vec{\mathbf{v}} &= \vec{\mathbf{v}}_D &&\text{on } \Gamma &&(14.160)
\end{aligned}$$

supplemented by

$$\int_\Omega p = 0$$

where $\boldsymbol{L}$ is the gradient tensor, $\mathbf{1}$ is the unit tensor.

Remark. *It may appear counter-intuitive at first to define $\boldsymbol{L}$ as being the gradient of the velocity instead of the strain rate tensor but under the assumption of incompressibility $\partial_x u + \partial_y v = 0$ (and constant viscosity) we can write:*

$$\vec{\nabla}\cdot(2\eta\boldsymbol{L}) = 2\eta\left(\begin{array}{c} \partial_x^2 u + \frac{1}{2}\partial_x\partial_y v + \frac{1}{2}\partial_y^2 u \\ \frac{1}{2}\partial_x^2 v + \frac{1}{2}\partial_y\partial_x u + \partial_y^2 v \end{array}\right) = 2\eta\left(\begin{array}{c} \partial_x^2 u + \frac{1}{2}\partial_x(-\partial_x u) + \frac{1}{2}\partial_y^2 u \\ \frac{1}{2}\partial_x^2 v + \frac{1}{2}\partial_y(-\partial_y v) + \partial_y^2 v \end{array}\right) = \eta\left(\begin{array}{c} \partial_x^2 u + \partial_y^2 u \\ \partial_x^2 v + \partial_y^2 v \end{array}\right) = \eta\Delta\vec{\mathbf{v}}$$

Remark. *Cockburn et al. (2010) [286] also introduce the vorticity-based formulation and the stress-based formulation but we will not explore these in what follows.*

RETYPE section 2.1 of [283]

Chapter 15

Tracking materials

Unless using a fully Lagrangian formulation, one needs an additional numerical method to represent/track the various materials present in an undeformable (Eulerian) mesh. The figure below (by B. Hillebrand) illustrates the three main methods used in geodynamics.
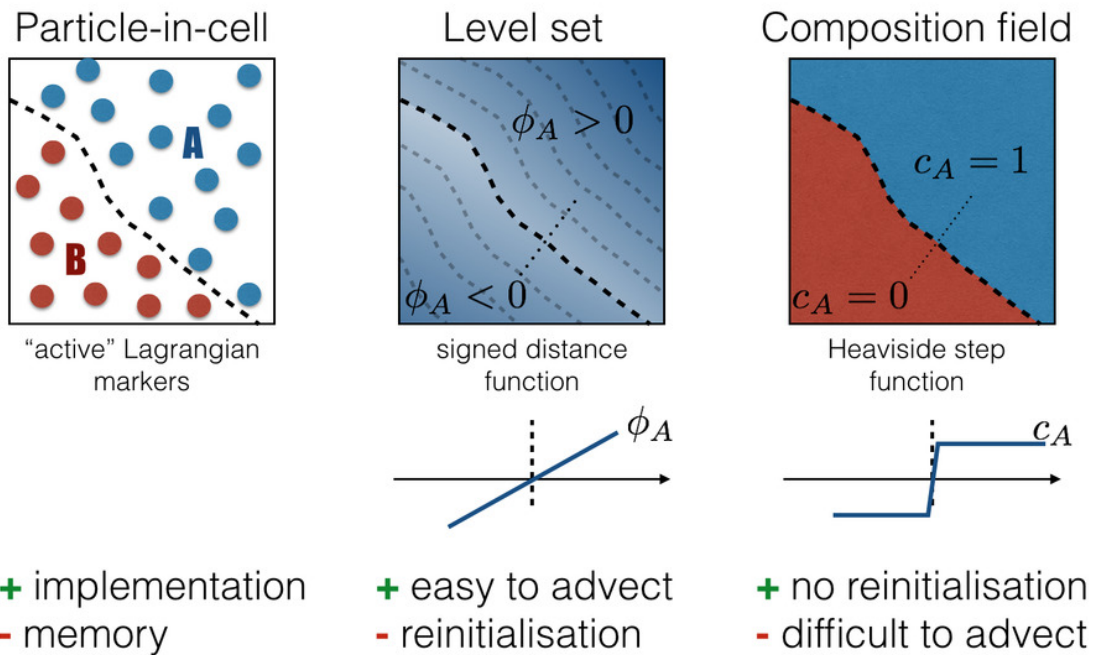
Note that what follows is applicable to FEM, FDM, etc ...

A typical test for advection algorithm is the Zalesak disk [1500]. It is a two dimensional test problem of solid body rotation with a constant angular velocity ω (in rad/sec):
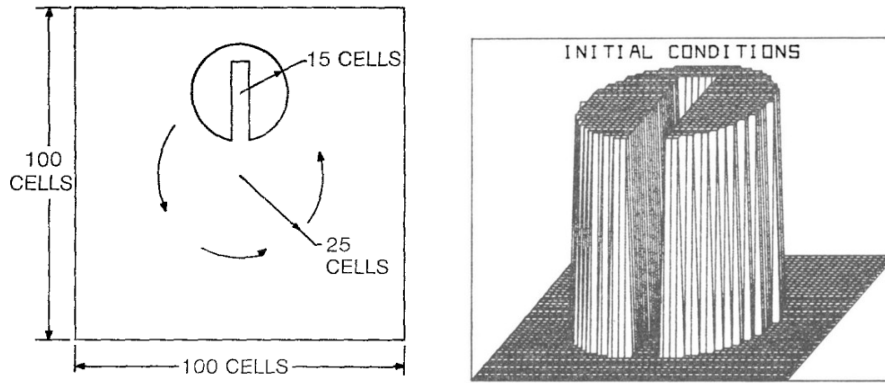
Taken from Zalesak [1500] (1979). Left: Schematic representation of two dimensional solid body rotation problem. The field inside the cut out has value 3 and it is 1 outside. The rotational speed is such that one full revolution is effected in 628 cycles. The width of the gap separating the two halves of the cylinder, as well as the maximum extent of the "bridge" connecting the two halves, is 5 cells. Right: Perspective view of initial conditions for the two dimensional! solid body rotation problem. Note that only a 50 × 50 portion of the mesh centered on the cylinder is displayed.

This benchmark is widely used in the literature, see for instance [1274, 1304, 1399, 1062, 59, 1516]. Note that the Zalesak disc is often supplemented with a cone and a Gaussian features:
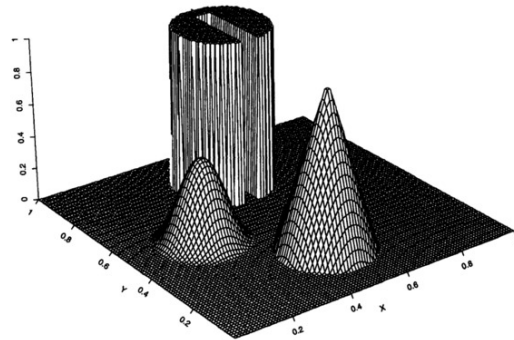
Taken from Leveque [823] (1996). Initial data for solid rotation tests

15.1 The Particle-in-cell technique

Remark. *The terms 'particle' and 'marker' are commonly (and unfortunately) interchangeably used in the literature in the context of the particle-in-cell technique. However, one should be aware that the marker-and-cell (MAC) technique is something different: it was invented in the early 60's at the Los Alamos Laboratories by Harlow and Welch [582] (1965). For more information on the MAC technique see the excellent review paper by McKee McKee et al. [900] (2008). Also, Tackley and King [1313] (2003) talk about the tracer-ratio method in the context of PIC...*

The Particle-in-cell method is by far the most widely used in computational geodynamics. In its most basic form it is a rather simple method to implement and this probably owes to its success and early adoption (e.g. Poliakov and Podlachikov [1073] (1992)) in non-parallel codes such as Sopale [452], I2VIS [481] or CitcomS [908]. It has been implemented in Aspect [464] and the inherent load balancing issues arising from the parallel implementation as well as from the use of Adaptive Mesh Refinement are discussed. It has also been implemented in the MILAMIN code [318] to study LLSVPs [972].

617
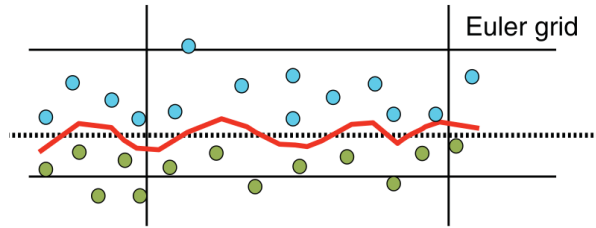
One of the main problems of the PIC method is the fact that the interface between the fluid is not tracked explicitely, and if one uses a random distribution of particles the black dotted line reprensents the 'real' interface between the fluids while the red line is liekly to be the interface one would obtain based on the distribution of particles. Taken from Crameri *et al.* (2012) [304].

Samuel [1178] (2018) does a great job at explaining the core problem with PIC:

The method requires the method requires particle-mesh and mesh-particle mappings to be specified. These critical operations constitute a major source of inaccuracy in the PIC solution [942, 375, 1340]. Indeed, while the Lagrangian advection alone is not prone to significant numerical diffusion, particle-mesh mappings can introduce important amounts of dissipation. This is particularly true when the spatial distribution of particles is not homogeneous, leading to areas in the vicinity of gridpoints that are not sufficiently well sampled by particles, and other regions where the domain is oversampled by particles. This recurrent sampling problem develops in regions characterized by strong deformation, and concerns both compressible and incompressible flow [1436, 1086]. The non-homogeneous sampling has two main origins.

- The first one corresponds to inaccuracies in advecting the Lagrangian particles [917]. This aspect has drawn the attention of a few recent studies [1436, 1086], which have proposed the use of conservative schemes to map velocity components from the Eulerian grid to the Lagrangian particles during their advection. Such schemes have shown to significantly improve the accuracy of the interpolation, and result in a considerably more homogeneous spatial sampling.

- The second origin, which has received less attention, is related to the deforming nature of the flow [951], and is completely independent of the accuracy of the numerical methods for interpolating the velocities at particles' locations. In fact, for a given velocity field, particles should travel along their characteristics, and even in the case of incompressible flows, the distance between characteristics can vary in general, and can strongly diverge or converge in regions characterized by strong deformation. This naturally leads to the development of a non-homogeneous spatial distribution of the Lagrangian particles, even if the particles locations are perfectly known.

A basic implementation of the PIC goes as follows:

1. distribute particles in the domain at startup,

2. assign a material identity (and/or any other quantity) to each particle,

3. project particle quantities on the nodes and/orelements of the mesh,

4. solve the Stokes equations for a new velocity field,

5. interpolate the velocity onto the particles,

6. move the particles with their respective velocities,

7. go back to step 3.

As it turns out each step above needs to be carefully executed and is more difficult than it first looks.

15.1.1 Distributing particles in the domain

Let us assume we wish to distribute N_p particles in the domain. How large must N_p be? To simplify, one end member could be 'as many particles as possible that fit in memory' while the other end member could be 'one per element/cell on average'. While the former does not necessarily guarantee a desired accuracy while being CPU and memory intensive, the latter will certainly lead to zones in the domain void of particles which will be problematic since the projection onto the mesh might yield zero values or very inaccurate values. How many particles (per element/cell) will be enough? Also, should the particles be randomly distributed in the domain or on some kind of regular grid? See STONE 13.

Taken from Tackley and King (2003) [1313]: "Tracers are initialized on a regular grid with each tracer perturbed from its grid position by a random amount of up to $\pm$ half a grid spacing, in order to eliminate artifacts due to tracer alignment."

15.1.2 Averaging and projection

This is a very critical step. Unfortunately, there is no community-wide agreed-upon method. The problem at hand boils down to: at a given location ($\vec{r}$) in space I need a the value of a field which is carried by the particles. The first step is to find the particle(s) close to this point. If done naively, this is a very costly affair, and begs the question what 'close' means. Finding all particles within a radius R of point $\vec{r}$ can be done very efficiently (e.g. with linked lists, Verlet lists, ...) but the choice of R proves to be critical: if too small, there may not be any particle inside the circle, and if too large there may be many particles inside the circle and the averaging over so many particles in space will prove to be over diffusive. In practice, the FD or FE mesh is used to provide an indication of R. In FDM, the four cells (or quarter cells) around a node represent the volume of space containing the particles whose properties are to be averaged [375] as illustrated in the following figure:
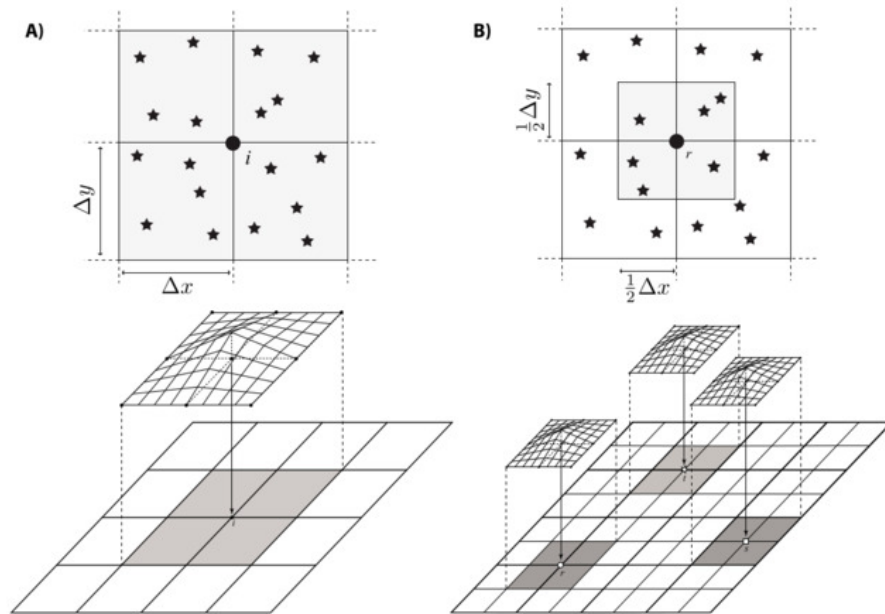
Taken from [375]. The "4-cell" and "1-cell" schemes for projecting properties defined on the markers (denoted by stars) onto a node (denoted by the solid circle). (A) The 4-cell scheme. The support of the interpolating function N_i associated with node i is indicated by the shaded region. Only markers within

Given that the FEM requires to compute integrals over each element, one could assume that only the particles inside the element will contribute to the average values assigned to the quadrature points (which I coin 'elemental approach').

However, one could also decide to first average the properties onto the nodes before using these nodal values to assign values to the quadrature points (which I coin 'nodal approach'). In this case the FDM approach seen above could apply.

Finally, in both FDM and FEM bi/trilinear basis functions are used for the interpolation as they can be interpreted as weighing functions. Higher order basis functions could also be used but the standard Q_2 basis functions (Section 8.3) are 2-nd order polynomials which can take negative values (as opposed to the Q_1 basis functions which are strictly positive) and this can pose problems: in some cases, although all values to be averaged are positive, their weighed average can be negative. See Section 16.26 for concrete examples.

nodal approach

elemental approach (1) - piece-wise constant interpolation

What follows is written with simplicity in mind, although more mathematical formulations can be found in the literature [464].

Assuming that we have established a list of particles tracking a field $f(\vec{r})$ inside the element we must now compute their average value $< f >$. The simplest approach which comes to mind is the arithmetic mean (am):

$$\langle f \rangle_{am} = \frac{\sum_{i=1}^{n} f_i}{n}$$

where n is the number of particles inside the element. In the case where f is the (mass) density ρ, it is indeed what should be used. However, turning now to viscosity η, we know that its value can vary by many orders of magnitude over very short distances. It is then likely that the average runs over values spanning values between 10^{18}Pa s and 10^{25}Pa s. As explained in [1200] the arithmetic averaging tends to 'favour' large values: if the sum runs over 10 particles, 9 carrying the value 10^{25} and 1 carrying the value 10^{19}, the average value is then

$$\langle \eta \rangle = \frac{9 \cdot 10^{25} + 1 \cdot 10^{19}}{10} \simeq 0.9 \cdot 10^{25}$$

which is much much closer to 10^{25} than to 10^{19}. Other averagings are then commonly used, namely the geometric mean (gm) and the harmonic mean (hm), defined as follows:

$$\langle f \rangle_{gm} = \left(\prod_i f_i \right)^{1/n} \qquad \text{or,} \qquad \log_{10}\langle f \rangle_{gm} = \frac{\sum_{i=1}^{n} \log_{10} f_i}{n}$$

and

$$\langle f \rangle_{hm} = \left(\frac{\sum_{i=1}^{n} \frac{1}{f_i}}{n} \right)^{-1} \qquad \text{or,} \qquad \frac{1}{\langle f \rangle_{hm}} = \frac{\sum_{i=1}^{n} \frac{1}{f_i}}{n}$$

The geometric mean can be seen as a form of arithmetic mean of $\log_{10}$ values, while the harmonic mean can be seen as a form of arithmetic mean of the inverse values.

Looking back at the above example, the geometric mean of the viscosities is given by

$$\log \langle \eta \rangle_{gm} = \frac{9 \cdot 25 + 1 \cdot 19}{10} = 24.4 \qquad \text{or,} \qquad \langle \eta \rangle_{gm} \simeq 2.5 \cdot 10^{24}$$

and the harmonic mean:

$$\langle \eta \rangle_{hm} \simeq \left(\frac{1}{10 \cdot 10^{19}} \right)^{-1} = 10^{20}$$

We see that the harmonic mean tends to favour the small values. Also we recover the known property:

$$\langle f \rangle_{am} \quad \geq \quad \langle f \rangle_{gm} \quad \geq \quad \langle f \rangle_{hm} \tag{15.1}$$

Once a single average value has been computed for the whole element, then all quadrature points are assigned this value.

elemental approach (2) - Least Squares Interpolation One can revisit this topic on the grounds that with high(er) order elements optimal convergence is unlikely to be reached if viscosity (and density) are assumed to be constant inside each element (see Gassmöller *et al.* (2019) [467]). One could therefore use the least-square method to arrive at a functional representation of the field inside the element which is as close as possible (in the least-squares sense, then) to the particle-based field.

Thielmann *et al.* (2014) [1340] use the $Q_2 P_{-1}$ element and introduce an element-wise interpolation scheme based on a least squares fitting of the particle properties and choose the functional to be a linear function to match the pressure space. They define the error ϵ such that

$$\epsilon^2 = \sum_{i=1}^{n} (\tilde{f}(x_i, y_i) - f_i)^2$$

with $\tilde{f}(x, y) = a + bx + cy$ and proceed to look for the minimum of ϵ^2, i.e. $\vec{\nabla}(\epsilon^2) = 0$ in the $\{a, b, c\}$ space:

$$
\begin{aligned}
0 = \frac{\partial \epsilon^2}{\partial a} &= 2 \sum_i (\tilde{f}(x_i, y_i) - f_i) \\
&= 2 \sum_i (a + bx_i + cy_i - f_i) \\
&= 2 \left[a \sum_i 1 + b \sum_i x_i + c \sum_i y_i - \sum_i f_i \right] \\
0 = \frac{\partial \epsilon^2}{\partial b} &= 2 \sum_i (\tilde{f}(x_i, y_i) - f_i) x_i \\
&= 2 \sum_i (a + bx_i + cy_i - f_i) x_i \\
&= 2 \left[a \sum_i x_i + b \sum_i x_i^2 + c \sum_i x_i y_i - \sum_i x_i f_i \right] \\
0 = \frac{\partial \epsilon^2}{\partial c} &= 2 \sum_i (\tilde{f}(x_i, y_i) - f_i) y_i \\
&= 2 \sum_i (a + bx_i + cy_i - f_i) y_i \\
&= 2 \left[a \sum_i y_i + b \sum_i x_i y_i + c \sum_i y_i^2 - \sum_i y_i f_i \right]
\end{aligned}
$$

so

$$
\begin{pmatrix}
\sum_i 1 & \sum_i x_i & \sum_i y_i \\
\sum_i x_i & \sum_i x_i^2 & \sum_i x_i y_i \\
\sum_i y_i & \sum_i x_i y_i & \sum_i y_i^2
\end{pmatrix}
\cdot
\begin{pmatrix} a \\ b \\ c \end{pmatrix}
=
\begin{pmatrix}
\sum_i f_i \\
\sum_i x_i f_i \\
\sum_i y_i f_i
\end{pmatrix}
$$

This method can trivially be extended to three dimensions. It must also be noted that it is not cheap: for each element the matrix and rhs above must be formed and the system solved for a, b, c.

We could also then decide to use a bi-linear function $\tilde{f}$, i.e.

$$\tilde{f}(x, y) = a + bx + cy + dxy$$

which lies in the Q_1 space of Taylor-Hood quadrilateral elements. In this case the error is

$$\epsilon^2 = \sum_{i=1}^{n} (\tilde{f}(x_i, y_i) - f_i)^2 = \sum_{i=1}^{n} (a + bx_i + cy_i + dx_i y_i - f_i)^2$$

and one has to solve a 4×4 system this time:

$$
\begin{pmatrix}
\sum_i 1 & \sum_i x_i & \sum_i y_i & \sum_i x_i y_i \\
\sum_i x_i & \sum_i x_i^2 & \sum_i x_i y_i & \sum_i x_i^2 y_i \\
\sum_i y_i & \sum_i x_i y_i & \sum_i y_i^2 & \sum_i x_i y_i^2 \\
\sum_i x_i y_i & \sum_i x_i y_i & \sum_i y_i^2 & \sum_i x_i^2 y_i^2
\end{pmatrix}
\cdot
\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}
=
\begin{pmatrix}
\sum_i f_i \\
\sum_i x_i f_i \\
\sum_i y_i f_i \\
\sum_i x_i y_i f_i
\end{pmatrix}
$$

which we write $\boldsymbol{A} \cdot \vec{c} = \boldsymbol{b}$. Note that the matrix $\boldsymbol{A}$ is symmetric. We see that this is a potentially numerically problematic equation. Distances/coordinates in geodynamic calculations are of the order of 100-1000km and viscosities are between 10^{19} and 10^{26}Pa s. The matrix would contain very large terms, which may compromise the accuracy of the system solve.

Once this linear system (or the previous one) has been solved we have obtained the coefficients $a, b, c(, d)$ which allow us to compute $\tilde{f}$ anywhere inside the element, and especially at the quadrature points. Once these coefficients have been obtained one can compute $\tilde{f}$ anywhere in the element, and in particular at the quadrature points.

Remark. *Using a different (bi)linear function $\tilde{f}$ for each element means that it is likely to be discontinuous from one element to another in regions of high gradients.*

There is however one drawback with this approach (linear or bi-linear alike): in the areas of steep gradients the computed coefficients can be such that the function $\tilde{f}$ evaluated on a quadrature point is negative which 1) would be wrong but not numerically dramatic for density, 2) would be wrong and physically and numerically problematic for viscosity (a viscosity cannot be negative, and this would automatically destroy the SPD nature of the viscous block of the Stokes matrix).

Least square fit of the density field for the sinking sphere experiment of Section 20.25.

Resolution is 33×33, 100 markers per element.
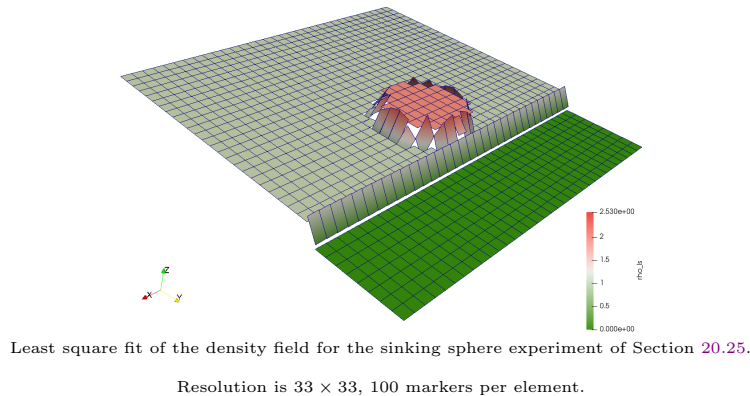
This problem is discussed in Thielmann *et al.* (2014) in Section 3.2.1 and they call this "Over- and Under-shooting". A simple (iterative) fix is then designed which insures that the computed value is within user-defined acceptable bounds. This is also mentioned in [467] but the authors explain that this problem was not encountered in the context of the publication.

Remark. *One could consider the above least-square approach with $\tilde{f} = a$, i.e. $\tilde{f}$ is a zero-th order polynomial. In this case*

$$\epsilon^2 = \sum_{i=1}^n (\tilde{f}(x_i, y_i) - f_i)^2 = \sum_{i=1}^n (a - f_i)^2$$

The gradient becomes

$$\vec{\nabla}(\epsilon^2) = \frac{d\epsilon^2}{da} = \sum_{i=1}^n 2(a - f_i) = 0$$

or $a = \frac{1}{n}\sum_i f_i$. We here recover the arithmetic averaging!

Remark. *Two variants of the PIC methods have been proposed: the Deformable PIC (DPIC) by Samuel (2018) [1178], and the multiscale PIC in [32].*

Remark. *TO BE WRITTEN. A word about the tracer ratio method. [1313]. Trim et al. (2020) show a modified method with a tracer repositioning algorithm designed to promote even tracer coverage [1371].*

Also look at Yang, Moresi, and Mansour [1479] and Bouffard, Labrosse, Choblet, Fournier, Aubert, and Tackley [127].

See STONE 67 for a concrete example of Particle-In-Cell use and a detailed explanation of its implementation. See also STONE 41 for an implementation of the least square method.

15.1.3   Interpolation of the velocity onto particles

.

Once the particle i has been localised inside a given element (Section 15.11) and its reduced coordinates (r, s, t) determined, the velocity at this location can be computed through the basis functions:

$$\vec{v}_i = \sum_{k=1}^m N_i(r, s, t)\vec{v}_k$$

This approach is not without problem: while the nodal velocities $\vec{v}_k$ are such that[1] $\vec{\nabla} \cdot \vec{v} = 0$ (in the weak sense), the computed velocity $\vec{v}_i$ is not necessarily divergence-free! In order to remedy this, a Conservative Velocity Interpolation (CVI) has been proposed in [1436]. Because the complete derivations for the CVI algorithm is quite large I have decided to make a new section about it (Section 15.2) rather than include it here.

15.1.4   Moving the particles

This is discussed in the context of the Runge-Kutta Methods, see Section 15.10.

Check Rene Gassmöller, Juliane Dannberg, Wolfgang Bangerth, Elbridge Gerry Puckett, and Cedric Thieulot. "Benchmarking the accuracy of higher order particle methods in geodynamic models of transient flow". In: *Geosci. Model. Dev.* 17 (2024), pp. 4115–4134. DOI: 10.5194/gmd-17-4115-2024

---

[1] for incompressible flows, of course

15.2 The Particle-in-cell technique - CVI style

To my knowledge the conservative velocity interpolation (CVI) was introduced to the computational geodynamics community in Wang, Agrusta, and van Hunen [1436] (2015). As mentioned in the paper "An improved velocity interpolation scheme that conserves the divergence of the flow field has been developed by Jenny, Pope, Muradoglu, and Caughey [682] (2001) and the simplified scheme for incompressible flow (i.e., divergence free) has been demonstrated that it largely eliminates the spurious distribution of particles for 2D incompressible flow problem (see Meyer and Jenny [917] (2004))."

Additional more recent publications on the topic of accurate marker advection: Sime, Maljaars, Wilson, and Keken [1246] (2021), Sime, Wilson, and Keken [1248] (2022).

15.2.1 A few remarks about Wang *et al.* (2015)

The article by Wang, Agrusta, and van Hunen [1436] (2015) comes with supplementary material with more details on the derivation of the corrective velocities but that material is a Word document printed to pdf with an annoying layout of equations, different font sizes, lack of alignment, etc ... Also, Fig. 1 of the paper is reproduced here:
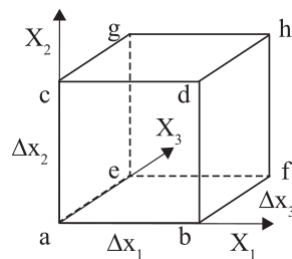
Why the authors chose to label nodes a,b,...h and not 1,2,...8 shall forever remain a mystery, but it is not as problematic as the labelling of the axes: indeed, if X_1 is the x-axis then X_3 should be the y-axis and X_2 the z-axis. That is quite illogical. Or is it a mistake in the drawing only? In any case this sheds some confusion on the equations presented in the paper so I have decided to carry out all the CVI derivations in this chapter.

Their paper does not seem to consider cases where the element is not a cuboid (so what about CitcomS, or ALE formulations?), nor does it address higher order elements. Finally many details of the setups in the paper are just not there and I had to email the author(s) multiple time regarding:

- the setup of the couette flow in section 3.1 is incomplete: for instance, size of the box ? velocity value ? exact formula for the vel field (couette flow, I know, but how thick are the layers before rotation)? etc ...
  Wang answered me: "The box is a unit box (nondimentional 1*1). I attached the function for the analytical solution for the exact formula for the velocity field that you asked. I didn't find the models file yet, so I can't tell you what it is the value of the velocity. But I think it can be: 1m*1m box with 1m/s on the surface (V0). In Citcom, the timestep is chosen to let any material in one cell not to move more than half of the cell length (CFL=0.5). Then we have this parameter "finetunedt" (< 1) to multiply it. I remember I usually use 0.9 or 0.7. So the CFL=0.45 or 0.35. Concerning the Couette flow we used a viscosity of 1e3, which make very sharp velocity contrast across the diagonal line."

```
for (i=1;i<=E->lmesh.nno;i++)
{
```

```
x =  E->X[1][i];
z =  E->X[2][i];
eta1=E->control.testvelval[1];
eta2=E->control.testvelval[2];
alpha=E->control.testvelval[3]*PI/180;  /*coordinate rotation angle */
V0=E->control.testvelval[4];
h=sqrt(2.0)*sin(alpha+PI/4); /*WHL: h (with analytical solution) is a function of the rotati
V1=(x*sin(alpha)+z*cos(alpha))*2*V0*eta2/(eta1+eta2)/h;
V2=(x*sin(alpha)+z*cos(alpha))*2*V0*eta1/(eta1+eta2)/h+(eta2-eta1)*V0/(eta1+eta2);
if (x*sin(alpha)+z*cos(alpha)<0.5*h)
{
E->V[1][i]=V1*cos(alpha);
E->V[2][i]=-V1*sin(alpha);
}
else
{
E->V[1][i]=V2*cos(alpha);
E->V[2][i]=-V2*sin(alpha);
}
if (E->mesh.nsd == 3)
E->V[3][i]=0.;
}
```

- which advection scheme was used and I am worried that at no point in the publication the timestep size is either mentioned nor its importance discussed.
  Wang answered: "About the timestep, my experience is that using smaller timestep would't solve this kind of problem. Otherwise we probably would not need to use this new velocity interpolation. I could not remember that I tested the effects of timestep for this model. So it would be nice to know the result if you test it. The advection scheme is the 2nd Runge Kutta. "

- Agrusta wrote: "here the input values for the couette flow: testvelval=100000,1,45,0.01 # eta1,eta2,angle,velocity. mesh = 33x33. initial tracers 100X100, random distribution"

Looking at their Fig. 2a,b we see black arrow tips in the blue region where velocity should be zero. Velocity is indeed zero and the authors confirmed that the arrow tips are an artefact of their visualisation software (!).

Relevant Literature: McNally (2011) [907] proposed a divergence-free interpolation of vector fields from point values in the context of magnetohydrodynamics. Pusok, Kaus, and Popov [1086] (2017) has applied the CVI to staggered grid FDM.

15.2.2 In 2D with Q_1 basis functions - Naive approach

Let us start directly in reduced coordinates $(r, s) \in [-1 : 1]^2$ (i.e. the reference element). The velocity components inside of the element are given by:

$$
\begin{aligned}
u^h(r, s) &= \sum_i \mathcal{N}_i(r, s) u_i \\
v^h(r, s) &= \sum_i \mathcal{N}_i(r, s) v_i
\end{aligned}
$$

where $\mathcal{N}_i$ are the four Q_1 basis functions defined as follows:

$$\mathcal{N}_1(r,s) = \frac{1}{4}(1-r)(1-s)$$

$$\mathcal{N}_2(r,s) = \frac{1}{4}(1+r)(1-s)$$

$$\mathcal{N}_3(r,s) = \frac{1}{4}(1+r)(1+s)$$

$$\mathcal{N}_4(r,s) = \frac{1}{4}(1-r)(1+s)$$

The incompressibility constraint in the $(r,s)-$coordinate system reads

$$(\vec{\nabla} \cdot \vec{v})^h = \frac{\partial u^h}{\partial r} + \frac{\partial v^h}{\partial s} = \sum_i \left(\frac{\partial \mathcal{N}_i}{\partial r} u_i + \frac{\partial \mathcal{N}_i}{\partial s} v_i \right) = 0.$$

However, it is trivial to verify that the incompressibility condition is not and *can not* be verified for all values of $r, s \in [-1, 1]^2$. It would then make sense to think of a corrective term to the interpolation which would add just enough degrees of freedoms so as to insure an exact[2] incompressibility in the element. Let us then write:

$$u^h(r,s) = \sum_i \mathcal{N}_i(r,s)u_i + (as+b)(1-r)(1+r)$$

$$v^h(r,s) = \sum_i \mathcal{N}_i(r,s)v_i + (cr+d)(1-s)(1+s)$$

Note that in this way the correction is zero on the $x = -1$ and $x = +1$ sides of the element for u, and likewise for v on the top and bottom sides (in other words the velocity remains continuous from one element to another). In this case,

$$\frac{\partial u^h}{\partial r} = \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (as+b)(-2r)$$

$$\frac{\partial v^h}{\partial s} = \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (cr+d)(-2s)$$

We have introduced 4 coefficients (a,b,c,d) which remain to be determined. We start with:

$$\sum_i \frac{\partial N_i}{\partial r} u_i = -\frac{1}{4}(1-s)u_1 + \frac{1}{4}(1-s)u_2 + \frac{1}{4}(1+s)u_3 - \frac{1}{4}(1+s)u_4$$

$$= (1-s)\frac{u_2 - u_1}{4} + (1+s)\frac{u_3 - u_4}{4}$$

$$= (1-s)u_{21} + (1+s)u_{34}$$

$$\sum_i \frac{\partial N_i}{\partial s} v_i = -\frac{1}{4}(1-r)v_1 - \frac{1}{4}(1+r)v_2 + \frac{1}{4}(1+r)v_3 + \frac{1}{4}(1-r)v_4$$

$$= (1-r)\frac{v_4 - v_1}{4} + (1+r)\frac{v_3 - v_2}{4}$$

$$= (1-r)v_{41} + (1+r)v_{32}$$

where $u_{ij} = (u_i - u_j)/4$ and $v_{ij} = (v_i - v_j)/4$, so that in the end

$$\frac{\partial u^h}{\partial r} = (1-s)u_{21} + (1+s)u_{34} + (as+b)(-2r) \tag{15.2}$$

$$\frac{\partial v^h}{\partial s} = (1-r)v_{41} + (1+r)v_{32} + (cr+d)(-2s) \tag{15.3}$$

---

[2]more on this later

626

The incompressibility condition is now:

$$(\vec{\nabla} \cdot \vec{v})^h = (1 - s)u_{21} + (1 + s)u_{34} + (as + b)(-2r) + (1 - r)v_{41} + (1 + r)v_{32} + (cr + d)(-2s) = 0$$

This can be rewritten as

$$(\vec{\nabla} \cdot \vec{v})^h = C_0 + C_1 r + C_2 s + C_3 rs = 0$$

where the four C_i coefficients are functions of the velocities and the other coefficients. In order for this expression to be exactly zero *everywhere*, each C coefficient has to be independently zero.

$$
\begin{array}{lll}
C_0 & (.) & u_{21} + u_{34} + v_{41} + v_{32} = 0 \\
C_1 & (r) & -v_{41} + v_{32} - 2b = 0 \\
C_2 & (s) & -u_{21} + u_{34} - 2d = 0 \\
C_3 & (rs) & -2a - 2c = 0
\end{array}
$$

The first line is simply the incompressibility condition expressed in the center of the element (i.e. $r = s = 0$), so we set it aside for now (I will come back to it later!) and focus on the remaining three.

At this stage it is important to note that in the absence of corrective terms (i.e. $a = b = c = d = 0$) then only $C_3 = 0$ and the divergence inside the element is a linear field.

We obtain

$$c = -a \qquad b = \frac{1}{2}(-v_{41} + v_{32}) \qquad d = \frac{1}{2}(-u_{21} + u_{34})$$

Since a and c are not otherwise constrained, we can set them to zero, and we then have:

$$b = \frac{1}{2}(v_{14} + v_{32}) \qquad d = \frac{1}{2}(u_{12} + u_{34})$$

and finally

$$
\begin{aligned}
u^h(r, s) &= \sum_i \mathcal{N}_i(r, s)u_i + b(1 - r)(1 + r) = \sum_i \mathcal{N}_i(r, s)u_i + \frac{1}{2}(v_{14} + v_{32})(1 - r)(1 + r) \\
v^h(r, s) &= \sum_i \mathcal{N}_i(r, s)v_i + d(1 - s)(1 + s) = \sum_i \mathcal{N}_i(r, s)v_i + \frac{1}{2}(u_{12} + u_{34})(1 - s)(1 + s)
\end{aligned}
$$

By using these corrected interpolations for both components of the velocity then one ensures that a point-wise divergence free velocity field anywhere in the element. However, these derivations were carried out in the reference element. In fact they would work also for rectangular elements with minimal changes, but not for generic quadrilaterals.

To be clear, let us now compute the velocity divergence of the corrected velocity field above:

$$
\begin{aligned}
(\vec{\nabla} \cdot \vec{v})^h &= \frac{\partial u^h}{\partial r} + \frac{\partial v^h}{\partial s} \\
&= (1 - s)u_{21} + (1 + s)u_{34} + \frac{1}{2}(v_{14} + v_{32})(-2r) + (1 - r)v_{41} + (1 + r)v_{32} + \frac{1}{2}(u_{12} + u_{34})(-2s) \\
&= u_{21} + u_{34} + v_{41} + v_{32} - su_{21} + su_{34} - rv_{14} - rv_{32} - rv_{41} + rv_{32} - su_{12} - su_{34} \\
&= u_{21} + u_{34} + v_{41} + v_{32} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15.4)
\end{aligned}
$$

A point must then be made crystal clear: the divergence is *not* zero. The quantity above is constant inside the element (it does not depend on r nor s). **All what the CVI algorithm does is to remove the spatial dependence of the velocity divergence inside the element**.

15.2.3 In 2D with Q_1 basis functions - better approach

We now consider a generic quadrilateral in the x, y-coordinate space and its equivalent in the reference space r, s. One can easily show that the gradient of a field f verifies

$$
\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \tilde{\boldsymbol{J}} \cdot \begin{pmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial s} \end{pmatrix}
$$

where $\tilde{\boldsymbol{J}}$ in the inverse of the Jacobian matrix. We then postulate again

$$
\begin{aligned}
u^h(r, s) &= \sum_i \mathcal{N}_i(r, s) u_i + (as + b)(1 - r)(1 + r) \\
v^h(r, s) &= \sum_i \mathcal{N}_i(r, s) v_i + (cr + d)(1 - s)(1 + s)
\end{aligned}
$$

In this case,

$$
\begin{aligned}
\frac{\partial u^h}{\partial r} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (as + b)(-2r) \\
\frac{\partial u^h}{\partial s} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial s} u_i + a(1 - r^2) \\
\frac{\partial v^h}{\partial r} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + c(1 - s^2) \\
\frac{\partial v^h}{\partial s} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (cr + d)(-2s)
\end{aligned}
$$

We have introduced 4 coefficients (a, b, c, d) which remain to be determined. In order to compute the velocity divergence inside the element we will need

$$
\begin{aligned}
\frac{\partial u}{\partial x} &= \tilde{J}_{xx} \frac{\partial u}{\partial r} + \tilde{J}_{xy} \frac{\partial u}{\partial s} \\
&= \tilde{J}_{xx} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (as + b)(-2r) \right) + \tilde{J}_{xy} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial s} u_i + a(1 - r^2) \right) \\
&= \tilde{J}_{xx} \left(-(1 - s)u_{12} + (1 + s)u_{34} + (as + b)(-2r) \right) \\
&+ \tilde{J}_{xy} \left(-(1 - r)u_{14} - (1 + r)u_{23} + a(1 - r^2) \right) \\
\frac{\partial v}{\partial y} &= \tilde{J}_{yx} \left(-(1 - s)v_{12} + (1 + s)v_{34} + c(1 - s^2) \right) \\
&+ \tilde{J}_{yy} \left(-(1 - r)v_{14} - (1 + r)v_{23} + (cr + d)(-2s) \right)
\end{aligned}
$$

where $u_{ij} = (u_i - u_j)/4$ and $v_{ij} = (v_i - v_j)/4$. The velocity divergence can be written as follows

$$
\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = C_0 + C_1 r + C_2 s + C_3 rs + C_4 r^2 + C_5 s^2 = 0
$$

with

$$
\begin{aligned}
C_0 &= J_{xx}(-u_{12} + u_{34}) + J_{xy}(-u_{14} - u_{23}) + J_{yx}(-v_{12} + v_{34}) + J_{yy}(-v_{14} - v_{23}) \\
C_1 &= J_{xy}(u_{14} - u_{23}) + J_{yy}(v_{14} - v_{23}) - 2bJ_{xx} \\
C_2 &= J_{xx}(u_{12} + u_{34}) + J_{yx}(v_{12} + v_{34}) - 2dJ_{yy} \\
C_3 &= -2aJ_{xx} - 2cJ_{yy} \\
C_4 &= -aJ_{xy} \\
C_5 &= -cJ_{yx}
\end{aligned}
$$

(15.5)

where the six C_i coefficients are functions of the velocities and the other coefficients. In order for this expression to be exactly null *everywhere*[3], each C coefficient has to be independently null.

This immediately yields $a = c = 0$ (since the components of the $\tilde{\boldsymbol{J}}$ tensor are not necessarily zero - and if J_{xy} and J_{yx} are zero then the equation for C_3 remains and we would still take $a = c = 0$ for simplicity) and the equation for C_3 is immediately satisfied. We then have:

$$
\begin{aligned}
b &= \frac{1}{2J_{xx}}(J_{xy}(u_{14} - u_{23}) + J_{yy}(v_{14} - v_{23})) \\
d &= \frac{1}{2J_{yy}}(J_{xx}(u_{12} + u_{34}) + J_{yx}(v_{12} + v_{34}))
\end{aligned}
$$

These expressions contain the same ingredients as before but also introduce more coupling between the velocity components. If the element is rectangular then $J_{xy} = J_{yx} = 0$ and

$$
\begin{aligned}
b &= \frac{J_{yy}}{2J_{xx}}(v_{14} - v_{23}) \\
d &= \frac{J_{xx}}{2J_{yy}}(u_{12} + u_{34})
\end{aligned}
$$

If the element is square then $J_{xx} = J_{yy} = 0$ so

$$
\begin{aligned}
b &= \frac{1}{2}(v_{14} - v_{23}) \\
d &= \frac{1}{2}(u_{12} + u_{34})
\end{aligned}
$$

and finally the velocity correction is

$$
\begin{aligned}
\delta u &= \frac{1}{2}(v_{14} - v_{23})(1 - r)(1 + r) \\
\delta v &= \frac{1}{2}(u_{12} + u_{34})(1 - s)(1 + s)
\end{aligned}
$$

(15.6)

15.2.4 Comparison with Wang *et al.* (2015) for 2D

Rather annoyingly Wang *et al.* (2015) use a reference element that is $[0, 1] \times [0, 1]$ as opposed to the standard $[-1, 1] \times [-1, 1]$:

---

[3]We know by now that this is not possible

Taken from the supplementary material of Wang *et al.* (2015).

Since basis functions must be 1 on their node, then the numbering must be as follows:

```
c--d            4--3
|  |     <=>    |  |
a--b            1--2
```

Setting $\Delta x_1 = \Delta x_2 = 1$, replacing a by 1, b by 2, c by 4 and d by 3, x_1 by r' and x_2 by s', U_1 by u and U_2 by v, we arrive at (in order to render the notations a bit lighter I have set $U = U_1$ and $V = U_2$)

$$\Delta U = \frac{1}{2}r'(1-r')(v_1 - v_2 - v_4 + v_3) = \frac{1}{2}r'(1-r')(4v_{14} - 4v_{23})$$

$$\Delta V = \frac{1}{2}s'(1-s')(u_1 - u_2 - u_4 + u_3) = \frac{1}{2}s'(1-s')(4u_{12} + 4u_{34})$$

Since $r = 2r' - 1$ and $s = 2s' - 1$ then we find that

$$\Delta U = \frac{1}{2}(1 - r^2)(v_{14} - v_{23})$$

$$\Delta V = \frac{1}{2}(1 - s^2)(u_{12} + u_{34}) \tag{15.7}$$

which is Eq. (15.6). In the case of the reference element then my velocity corrections are identical to theirs.

Let us look at the equations of the figure above. Since the authors state that they "transform the rectangular cells into unit squares" we do away with $\Delta x_1 = \Delta x_2 = 1$. Eqs. 3 and 1 together yield:

$$U = (1-x_1)(1-x_2)U^a + x_1(1-x_2)U^b + (1-x_1)x_2 U^c + x_1 x_2 U^d + \frac{1}{2}x_1(1-x_1)(V^a - V^b - V^c + V^d)$$

$$V = (1-x_1)(1-x_2)V^a + x_1(1-x_2)V^b + (1-x_1)x_2 V^c + x_1 x_2 V^d + \frac{1}{2}x_2(1-x_2)(U^a - U^b - U^c + U^d)$$

Then

$$\frac{\partial U}{\partial x_1} = -(1-x_2)U^a + (1-x_2)U^b - x_2 U^c + x_2 U^d + \frac{1}{2}(1-2x_1)(V^a - V^b - V^c + V^d)$$

$$\frac{\partial V}{\partial x_2} = -(1-x_1)V^a - x_1 V^b + (1-x_1)V^c + x_1 V^d + \frac{1}{2}(1-2x_2)(U^a - U^b - U^c + U^d)$$

So

$$
\begin{aligned}
\frac{\partial U}{\partial x_1} + \frac{\partial V}{\partial x_2} = \ & -(1 - x_2)U^a + (1 - x_2)U^b - x_2 U^c + x_2 U^d + \frac{1}{2}(1 - 2x_1)(V^a - V^b - V^c + V^d) \\
& -(1 - x_1)V^a - x_1 V^b + (1 - x_1)V^c + x_1 V^d + \frac{1}{2}(1 - 2x_2)(U^a - U^b - U^c + U^d) \\
= \ & -U^a + U^b + x_2(U^a - U^b - U^c + U^d) + \frac{1}{2}(V^a - V^b - V^c + V^d) - x_1(V^a - V^b - V^c + V^d) \\
& -V^a + V^c + x_1(V^a - V^b - V^c + V^d) + \frac{1}{2}(U^a - U^b - U^c + U^d) - x_2(U^a - U^b - U^c + U^d) \\
= \ & -U^a + U^b + \frac{1}{2}(V^a - V^b - V^c + V^d) - V^a + V^c + \frac{1}{2}(U^a - U^b - U^c + U^d) \\
\neq \ & 0 \tag{15.8}
\end{aligned}
$$

Unfortunately, the authors seem to be under the impression that this quantity is zero since they talk of "2D divergence-free interpolation" and "the divergence of the vector field need to be zero". Their own equations prove that this is not the case.

15.2.5 In 3D with Q_1 basis functions - Naive approach

In this case we are addressing the case of the divergence being as close to zero as possible in the reference element. We'll treat the case of a generic hexahedron in the next section.

Let us start directly in reduced coordinates $(r, s, t) \in [-1 : 1]^3$:

$$
\begin{aligned}
u^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) u_i \\
v^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) v_i \\
w^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) w_i
\end{aligned}
$$

with

$$
\begin{aligned}
\mathcal{N}_1 &= \frac{1}{8}(1 - r)(1 - s)(1 - t) \\
\mathcal{N}_2 &= \frac{1}{8}(1 + r)(1 - s)(1 - t) \\
\mathcal{N}_3 &= \frac{1}{8}(1 + r)(1 + s)(1 - t) \\
\mathcal{N}_4 &= \frac{1}{8}(1 - r)(1 + s)(1 - t) \\
\mathcal{N}_5 &= \frac{1}{8}(1 - r)(1 - s)(1 + t) \\
\mathcal{N}_6 &= \frac{1}{8}(1 + r)(1 - s)(1 + t) \\
\mathcal{N}_7 &= \frac{1}{8}(1 + r)(1 + s)(1 + t) \\
\mathcal{N}_8 &= \frac{1}{8}(1 - r)(1 + s)(1 + t)
\end{aligned}
$$

The incompressibility constraint imposes:

$$
\frac{\partial u^h}{\partial r} + \frac{\partial v^h}{\partial s} + \frac{\partial w^h}{\partial t} = 0 = \sum_i \left(\frac{\partial \mathcal{N}_i}{\partial r} u_i + \frac{\partial \mathcal{N}_i}{\partial s} v_i + \frac{\partial \mathcal{N}_i}{\partial t} w_i \right) = 0
$$

However, once again it is trivial to verify that the incompressibility condition is not and can not be verified for all values of $r, s, t \in [-1, 1]^3$.

It would then make sense to think of a corrective term to the interpolation which would add just enough degrees of freedoms so as to insure an exact incompressibility in the element. Let us then write:

$$
\begin{aligned}
u^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) u_i + (as + bt + c)(1 - r)(1 + r) \\
v^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) v_i + (dr + et + f)(1 - s)(1 + s) \\
w^h(r, s, t) &= \sum_i \mathcal{N}_i(r, s, t) w_i + (gr + hs + i)(1 - t)(1 + t)
\end{aligned}
$$

We thereby make sure that the corrections are zero on the edges so that velocity remains continuous from one element to another. In this case,

$$
\begin{aligned}
\frac{\partial u^h}{\partial r} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (as + bt + c)(-2r) \\
\frac{\partial v^h}{\partial s} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (dr + et + f)(-2s) \\
\frac{\partial w^h}{\partial t} &= \sum_i \frac{\partial \mathcal{N}_i}{\partial t} w_i + (gr + hs + i)(-2t)
\end{aligned}
$$

We have introduced 9 coefficients $(a, b, c, d, e, f, g, h, i)$ which remain to be determined. The incompressibility condition is now:

$$
\sum_i \left(\frac{\partial \mathcal{N}_i}{\partial r} u_i + \frac{\partial \mathcal{N}_i}{\partial s} v_i + \frac{\partial \mathcal{N}_i}{\partial t} w_i \right) + (as + bt + c)(-2r) + (dr + et + f)(-2s) + (gr + hs + i)(-2t) = 0
$$

This can be rewritten as

$$
C_0 + C_1 r + C_2 s + C_3 t + C_4 rs + C_5 st + C_6 rt = 0
$$

where the seven C_i coefficients are functions of the velocities and the other coefficients. In order for this expression to be exactly zero *everywhere*[4], each C coefficient has to be independently zero.

We start with:

$$
\begin{aligned}
8 \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i &= (1 - s)(1 - t)(u_2 - u_1) + (1 + s)(1 - t)(u_3 - u_4) + (1 - s)(1 + t)(u_6 - u_5) + (1 + s)(1 + t)(\\
8 \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i &= (1 - r)(1 - t)(v_4 - v_1) + (1 + r)(1 - t)(v_3 - v_2) + (1 - r)(1 + t)(v_8 - v_5) + (1 + r)(1 + t)(v \\
8 \sum_i \frac{\partial \mathcal{N}_i}{\partial t} w_i &= (1 - r)(1 - s)(w_5 - w_1) + (1 + r)(1 - s)(w_6 - w_2) + (1 + r)(1 + s)(w_7 - w_3) + (1 - r)(1 +
\end{aligned}
$$

Let us denote $u_{ij} = (u_i - v_j)/8$ (same for v, w), so that:

$$
\begin{aligned}
\sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i &= (1 - s)(1 - t) u_{21} + (1 + s)(1 - t) u_{34} + (1 - s)(1 + t) u_{65} + (1 + s)(1 + t) u_{78} \\
\sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i &= (1 - r)(1 - t) v_{41} + (1 + r)(1 - t) v_{32} + (1 - r)(1 + t) v_{85} + (1 + r)(1 + t) v_{76} \\
\sum_i \frac{\partial \mathcal{N}_i}{\partial t} w_i &= (1 - r)(1 - s) w_{51} + (1 + r)(1 - s) w_{62} + (1 + r)(1 + s) w_{73} + (1 - r)(1 + s) w_{84}
\end{aligned}
$$

---

[4]By now we know this is not possible – see 2D

632

We finally arrive at:

$$
\begin{aligned}
C_0 \quad (.) \quad & u_{21} + u_{34} + u_{65} + u_{78} + v_{41} + v_{32} + v_{85} + v_{76} + w_{51} + w_{62} + w_{73} + w_{84} = 0 \\
C_1 \quad (r) \quad & -v_{41} + v_{32} - v_{85} + v_{76} - w_{51} + w_{62} + w_{73} - w_{84} - 2c = 0 \\
C_2 \quad (s) \quad & -u_{21} + u_{34} - u_{65} + u_{78} - w_{51} - w_{62} + w_{73} + w_{84} - 2f = 0 \\
C_3 \quad (t) \quad & -u_{21} - u_{34} + u_{65} + u_{78} - v_{41} - v_{32} + v_{85} + v_{76} - 2i = 0 \\
C_4 \quad (rs) \quad & w_{51} - w_{62} + w_{73} - w_{84} - 2a - 2d = 0 \\
C_5 \quad (st) \quad & u_{21} - u_{34} - u_{65} + u_{78} - 2e - 2h = 0 \\
C_6 \quad (rt) \quad & v_{41} - v_{32} - v_{85} + v_{76} - 2b - 2g = 0
\end{aligned}
$$

I leave C_0 alone but I still unfortunately end up with 6 equations and 9 unknowns a, b, c, d, e, f, g, h. Coming up with additional constraints is not trivial, so I will instead further assume $\alpha_r = b = a$, $\alpha_s = e = d$ and $\alpha_t = h = g$, and rename $\beta_r = c$, $\beta_s = f$ and $\beta_t = i$ so that I have now six unknowns $\alpha_r, \alpha_s, \alpha_t, \beta_r, \beta_s, \beta_t$ for six equations

$$
\begin{aligned}
C_1 \quad (r) \quad & -v_{41} + v_{32} - v_{85} + v_{76} - w_{51} + w_{62} + w_{73} - w_{84} - 2\beta_r \\
C_2 \quad (s) \quad & -u_{21} + u_{34} - u_{65} + u_{78} - w_{51} - w_{62} + w_{73} + w_{84} - 2\beta_s \\
C_3 \quad (t) \quad & -u_{21} - u_{34} + u_{65} + u_{78} - v_{41} - v_{32} + v_{85} + v_{76} - 2\beta_t \\
C_4 \quad (rs) \quad & w_{51} - w_{62} + w_{73} - w_{84} - 2\alpha_r - 2\alpha_s \\
C_5 \quad (st) \quad & u_{21} - u_{34} - u_{65} + u_{78} - 2\alpha_s - 2\alpha_t \\
C_6 \quad (rt) \quad & v_{41} - v_{32} - v_{85} + v_{76} - 2\alpha_r - 2\alpha_t
\end{aligned}
$$

This naturally yields:

$$
\begin{aligned}
\beta_r \; &= \; \frac{1}{2}(-v_{41} + v_{32} - v_{85} + v_{76} - w_{51} + w_{62} + w_{73} - w_{84}) \\
&= \; \frac{1}{16}(v_1 - v_2 + v_3 - v_4 + v_5 - v_6 + v_7 - v_8 + w_1 - w_2 - w_3 + w_4 - w_5 + w_6 + w_7 - w_8) \\
\beta_s \; &= \; \frac{1}{2}(-u_{21} + u_{34} - u_{65} + u_{78} - w_{51} - w_{62} + w_{73} + w_{84}) \\
&= \; \frac{1}{16}(u_1 - u_2 + u_3 - u_4 + u_5 - u_6 + u_7 - u_8 + w_1 + w_2 - w_3 - w_4 - w_5 - w_6 + w_7 + w_8) \\
\beta_t \; &= \; \frac{1}{2}(-u_{21} - u_{34} + u_{65} + u_{78} - v_{41} - v_{32} + v_{85} + v_{76}) \\
&= \; \frac{1}{16}(u_1 - u_2 - u_3 + u_4 - u_5 + u_6 + u_7 - u_8 + v_1 + v_2 - v_3 - v_4 - v_5 - v_6 + v_7 + v_8)
\end{aligned}
$$

and we need to solve

$$
\begin{aligned}
\tilde{w} - 2\alpha_r - 2\alpha_s \; &= \; 0 \\
\tilde{u} - 2\alpha_s - 2\alpha_t \; &= \; 0 \\
\tilde{v} - 2\alpha_r - 2\alpha_t \; &= \; 0
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{u} \; &= \; u_{21} - u_{34} - u_{65} + u_{78} = \frac{1}{8}(-u_1 + u_2 - u_3 + u_4 + u_5 - u_6 + u_7 - u_8) \\
\tilde{v} \; &= \; v_{41} - v_{32} - v_{85} + v_{76} = \frac{1}{8}(-v_1 + v_2 - v_3 + v_4 + v_5 - v_6 + v_7 - v_8) \\
\tilde{w} \; &= \; w_{51} - w_{62} + w_{73} - w_{84} = \frac{1}{8}(-w_1 + w_2 - w_3 + w_4 + w_5 - w_6 + w_7 - w_8)
\end{aligned}
$$

which yields:

$$\alpha_r = \frac{1}{4}(-\tilde{u} + \tilde{v} + \tilde{w}) \qquad \alpha_s = \frac{1}{4}(\tilde{u} - \tilde{v} + \tilde{w}) \qquad \alpha_t = \frac{1}{4}(\tilde{u} + \tilde{v} - \tilde{w})$$

So finally:

$$
\begin{aligned}
u^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)u_i + [\alpha_r(s+t) + \beta_r](1-r)(1+r) \\
v^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)v_i + [\alpha_s(r+t) + \beta_s](1-s)(1+s) \\
w^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)w_i + [\alpha_t(r+s) + \beta_t](1-t)(1+t)
\end{aligned}
$$

15.2.6 In 3D with Q_1 basis functions - better approach

We start again from

$$
\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix} = \tilde{\boldsymbol{J}} \cdot \begin{pmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial s} \\ \frac{\partial f}{\partial t} \end{pmatrix}
$$

where $\tilde{\boldsymbol{J}}$ is the inverse of the Jacobian matrix $\boldsymbol{J}$. We then postulate

$$
\begin{aligned}
u^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)u_i + (as + bt + c)(1-r)(1+r) \\
v^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)v_i + (dr + et + f)(1-s)(1+s) \\
w^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)w_i + (gr + hs + i)(1-t)(1+t)
\end{aligned}
$$

so that:

$$
\begin{aligned}
\frac{\partial u}{\partial x} &= \tilde{J}_{xx}\frac{\partial u^h}{\partial r} + \tilde{J}_{xy}\frac{\partial u}{\partial s} + \tilde{J}_{xz}\frac{\partial u}{\partial t} \\
&= \tilde{J}_{xx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}u_i + (as+bt+c)(-2r)\right] + \tilde{J}_{xy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}u_i + a(1-r^2)\right] + \tilde{J}_{xz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}u_i + b(1-r^2)\right] \\
\frac{\partial v}{\partial y} &= \tilde{J}_{yx}\frac{\partial v^h}{\partial r} + \tilde{J}_{yy}\frac{\partial v}{\partial s} + \tilde{J}_{yz}\frac{\partial v}{\partial t} \\
&= \tilde{J}_{yx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}v_i + d(1-s^2)\right] + \tilde{J}_{yy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}v_i + (dr+et+f)(-2s)\right] + \tilde{J}_{yz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}v_i + e(1-s^2)\right] \\
\frac{\partial w}{\partial z} &= \tilde{J}_{zx}\frac{\partial w^h}{\partial r} + \tilde{J}_{zy}\frac{\partial w}{\partial s} + \tilde{J}_{zz}\frac{\partial w}{\partial t} \\
&= \tilde{J}_{zx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}w_i + g(1-t^2)\right] + \tilde{J}_{zy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}w_i + h(1-t^2)\right] + \tilde{J}_{zz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}w_i + (gr+hs+i)(-2t)\right]
\end{aligned}
$$

where for any function f:

$$\sum_i \frac{\partial \mathcal{N}_i}{\partial r} f_i = (1-s)(1-t)f_{21} + (1-s)(1+t)f_{65} + (1+s)(1-t)f_{34} + (1+s)(1+t)f_{78}$$

$$\begin{aligned}
&= (f_{21} + f_{65} + f_{34} + f_{78}) \\
&+ (-f_{21} - f_{65} + f_{34} + f_{78})s \\
&+ (-f_{21} + f_{65} - f_{34} + f_{78})t \\
&+ (f_{21} - f_{65} - f_{34} + f_{78})st \\
&= f_{r1} + f_{r2}s + f_{r3}t + f_{r4}st
\end{aligned}$$

$$\sum_i \frac{\partial \mathcal{N}_i}{\partial s} f_i = (1-r)(1-t)f_{41} + (1+r)(1-t)f_{32} + (1-r)(1+t)f_{85} + (1+r)(1+t)f_{76}$$

$$\begin{aligned}
&= (f_{41} + f_{32} + f_{85} + f_{76}) \\
&+ (-f_{41} + f_{32} - f_{85} + f_{76})r \\
&+ (-f_{41} - f_{32} + f_{85} + f_{76})t \\
&+ (f_{41} - f_{32} - f_{85} + f_{76})rt \\
&= f_{s1} + f_{s2}r + f_{s3}t + f_{s4}rt
\end{aligned}$$

$$\sum_i \frac{\partial \mathcal{N}_i}{\partial t} f_i = (1-r)(1-s)f_{51} + (1+r)(1-s)f_{62} + (1+r)(1+s)f_{73} + (1-r)(1+s)f_{84}$$

$$\begin{aligned}
&= (f_{51} + f_{62} + f_{73} + f_{84}) \\
&+ (-f_{51} + f_{62} + f_{73} - f_{84})r \\
&+ (-f_{51} - f_{62} + f_{73} + f_{84})s \\
&+ (f_{51} - f_{62} + f_{73} - f_{84})rs \\
&= f_{t1} + f_{t2}r + f_{t3}s + f_{t4}rs
\end{aligned}$$

The velocity divergence is then

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$$

$$= \tilde{J}_{xx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}u_i + (as+bt+c)(-2r)\right] + \tilde{J}_{xy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}u_i + a(1-r^2)\right] + \tilde{J}_{xz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}u_i + b(1-r^2)\right]$$

$$+ \tilde{J}_{yx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}v_i + d(1-s^2)\right] + \tilde{J}_{yy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}v_i + (dr+et+f)(-2s)\right] + \tilde{J}_{yz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}v_i + e(1-s^2)\right]$$

$$+ \tilde{J}_{zx}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial r}w_i + g(1-t^2)\right] + \tilde{J}_{zy}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial s}w_i + h(1-t^2)\right] + \tilde{J}_{zz}\left[\sum_i \frac{\partial \mathcal{N}_i}{\partial t}w_i + (gr+hs+i)(-2t)\right]$$

$$= \tilde{J}_{xx}\left[u_{r1} + u_{r2}s + u_{r3}t + u_{r4}st + (as+bt+c)(-2r)\right]$$
$$+ \tilde{J}_{xy}\left[u_{s1} + u_{s2}r + u_{s3}t + u_{s4}rt + a(1-r^2)\right]$$
$$+ \tilde{J}_{xz}\left[u_{t1} + u_{t2}r + u_{t3}s + u_{t4}rs + b(1-r^2)\right]$$
$$+ \tilde{J}_{yx}\left[v_{r1} + v_{r2}s + v_{r3}t + v_{r4}st + d(1-s^2)\right]$$
$$+ \tilde{J}_{yy}\left[v_{s1} + v_{s2}r + v_{s3}t + v_{s4}rt + (dr+et+f)(-2s)\right]$$
$$+ \tilde{J}_{yz}\left[v_{t1} + v_{t2}r + v_{t3}s + v_{t4}rs + e(1-s^2)\right]$$
$$+ \tilde{J}_{zx}\left[w_{r1} + w_{r2}s + w_{r3}t + w_{r4}st + g(1-t^2)\right]$$
$$+ \tilde{J}_{zy}\left[w_{s1} + w_{s2}r + w_{s3}t + w_{s4}rt + h(1-t^2)\right]$$
$$+ \tilde{J}_{zz}\left[w_{t1} + w_{t2}r + w_{t3}s + w_{t4}rs + (gr+hs+i)(-2t)\right]$$
$$= C_0 + C_1 r + C_2 s + C_3 t + C_4 rs + C_5 st + C_6 rt + C_7 r^2 + C_8 s^2 + C_9 t^2 = 0 \qquad (15.9)$$

with:

$$C_0 = \tilde{J}_{xx}u_{r1} + \tilde{J}_{xy}u_{s1} + \tilde{J}_{xz}u_{t1} + \tilde{J}_{yx}v_{r1} + \tilde{J}_{yy}v_{s1} + \tilde{J}_{yz}v_{t1} + \tilde{J}_{zx}w_{r1} + \tilde{J}_{zy}w_{s1} + \tilde{J}_{zz}w_{t1}$$
$$+ \tilde{J}_{xy}a + \tilde{J}_{xz}b + \tilde{J}_{yx}d + \tilde{J}_{yz}e + \tilde{J}_{zx}g + \tilde{J}_{zy}h$$
$$C_1 = \tilde{J}_{xy}u_{s2} + \tilde{J}_{xz}u_{t2} + \tilde{J}_{yy}v_{s2} + \tilde{J}_{yz}v_{t2} + \tilde{J}_{zy}w_{s2} + \tilde{J}_{zz}w_{t2} - \tilde{J}_{xx}2c$$
$$C_2 = \tilde{J}_{xx}u_{r2} + \tilde{J}_{xz}u_{t3} + \tilde{J}_{yx}v_{r2} + \tilde{J}_{yz}v_{t3} + \tilde{J}_{zx}w_{r2} + \tilde{J}_{zz}w_{t3} - \tilde{J}_{yy}2f$$
$$C_3 = \tilde{J}_{xx}u_{r3} + \tilde{J}_{xy}u_{s3} + \tilde{J}_{yx}v_{r3} + \tilde{J}_{yy}v_{s3} + \tilde{J}_{zx}w_{r3} + \tilde{J}_{zy}w_{s3} - \tilde{J}_{zz}2i$$
$$C_4 = \tilde{J}_{xz}u_{t4} + \tilde{J}_{yz}v_{t4} + \tilde{J}_{zz}w_{t4} - \tilde{J}_{xx}2a - \tilde{J}_{yy}2d$$
$$C_5 = \tilde{J}_{xx}u_{r4} + \tilde{J}_{yx}v_{r4} + \tilde{J}_{zx}w_{r4} - \tilde{J}_{yy}2e - \tilde{J}_{zz}2h$$
$$C_6 = \tilde{J}_{xy}u_{s4} + \tilde{J}_{yy}v_{s4} + \tilde{J}_{zy}w_{s4} - \tilde{J}_{xx}2b - \tilde{J}_{zz}2g$$
$$C_7 = -\tilde{J}_{xy}a - \tilde{J}_{xz}b$$
$$C_8 = -\tilde{J}_{yx}d - \tilde{J}_{yz}e$$
$$C_9 = -\tilde{J}_{zx}g - \tilde{J}_{zy}h$$

Of course what we want is a point-wise zero velocity divergence so we would need $C_0 = C_1 = ...C_9 = 0$. However we have 10 C coefficients/equations and only 9 variables $a, b, c, d, e, f, g, h, i$. We leave the C_0 equation alone and hope for the best (see 2D case). In other words we hope that if/when we have found $a, b, c, d, e, f, g, h, i$ so that $C_1 = ...C_9 = 0$ then C_0 is 'small' (whatever that means). As mentioned earlier, the CVI only removes the spatial dependence of the velocity divergence inside an element, it does not zero it.

It is then trivial to obtain c, f, i from the equations of C_1, C_2, C_3:

$$C_1 = 0 \;\Rightarrow\; \tilde{J}_{xy}u_{s2} + \tilde{J}_{xz}u_{t2} + \tilde{J}_{yy}v_{s2} + \tilde{J}_{yz}v_{t2} + \tilde{J}_{zy}w_{s2} + \tilde{J}_{zz}w_{t2} - \tilde{J}_{xx}2c = 0$$

$$c = \frac{1}{2\tilde{J}_{xx}}(\tilde{J}_{xy}u_{s2} + \tilde{J}_{xz}u_{t2} + \tilde{J}_{yy}v_{s2} + \tilde{J}_{yz}v_{t2} + \tilde{J}_{zy}w_{s2} + \tilde{J}_{zz}w_{t2})$$

$$C_2 = 0 \;\Rightarrow\; \tilde{J}_{xx}u_{r2} + \tilde{J}_{xz}u_{t3} + \tilde{J}_{yx}v_{r2} + \tilde{J}_{yz}v_{t3} + \tilde{J}_{zx}w_{r2} + \tilde{J}_{zz}w_{t3} - \tilde{J}_{yy}2f = 0$$

$$f = \frac{1}{2\tilde{J}_{yy}}(\tilde{J}_{xx}u_{r2} + \tilde{J}_{xz}u_{t3} + \tilde{J}_{yx}v_{r2} + \tilde{J}_{yz}v_{t3} + \tilde{J}_{zx}w_{r2} + \tilde{J}_{zz}w_{t3})$$

$$C_3 = 0 \;\Rightarrow\; \tilde{J}_{xx}u_{r3} + \tilde{J}_{xy}u_{s3} + \tilde{J}_{yx}v_{r3} + \tilde{J}_{yy}v_{s3} + \tilde{J}_{zx}w_{r3} + \tilde{J}_{zy}w_{s3} - \tilde{J}_{zz}2i = 0$$

$$i = \frac{1}{2\tilde{J}_{zz}}(\tilde{J}_{xx}u_{r3} + \tilde{J}_{xy}u_{s3} + \tilde{J}_{yx}v_{r3} + \tilde{J}_{yy}v_{s3} + \tilde{J}_{zx}w_{r3} + \tilde{J}_{zy}w_{s3})$$

Concerning a, b, d, e, g, h we are left with 6 equations for 6 unknowns, which can be cast as follows:

$$\begin{pmatrix} \tilde{J}_{xx} & & \tilde{J}_{yy} & & & \\ & & \tilde{J}_{yy} & & \tilde{J}_{zz} & \\ & \tilde{J}_{xx} & & & \tilde{J}_{zz} & \\ \tilde{J}_{xy} & \tilde{J}_{xz} & & & & \\ & & \tilde{J}_{yx} & \tilde{J}_{yz} & & \\ & & & & \tilde{J}_{zx} & \tilde{J}_{zy} \end{pmatrix} \begin{pmatrix} a \\ b \\ d \\ e \\ g \\ h \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \tilde{J}_{xz}u_{t4} + \tilde{J}_{yz}v_{t4} + \tilde{J}_{zz}w_{t4} \\ \tilde{J}_{xx}u_{r4} + \tilde{J}_{yx}v_{r4} + \tilde{J}_{zx}w_{r4} \\ \tilde{J}_{xy}u_{s4} + \tilde{J}_{yy}v_{s4} + \tilde{J}_{zy}w_{s4} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

At this stage we can only hope that the system is not ill-posed and that a solution exists. Obviously solving a 6×6 linear system for every marker/particle/etc ... will turn out to be costly. Let's see if we cannot do better.

From the last three equations for C_7, C_8, C_9 we have

$$b = -\frac{\tilde{J}_{xy}}{\tilde{J}_{xz}}a \qquad d = -\frac{\tilde{J}_{yz}}{\tilde{J}_{yx}}e \qquad h = -\frac{\tilde{J}_{zx}}{\tilde{J}_{zy}}g$$

At this stage we have determined c, f, i entirely and have expressed b, d, h as functions of a, e, g. There only remain three unknowns a, e, g and the equations involving C_4, C_5, C_6 become:

$$0 = C_4 = \underbrace{\tilde{J}_{xz}u_{t4} + \tilde{J}_{yz}v_{t4} + \tilde{J}_{zz}w_{t4}}_{2T} - \tilde{J}_{xx}2a - \tilde{J}_{yy}2d = 2T - \tilde{J}_{xx}2a + \tilde{J}_{yy}2\frac{\tilde{J}_{yz}}{\tilde{J}_{yx}}e$$

$$0 = C_5 = \underbrace{\tilde{J}_{xx}u_{r4} + \tilde{J}_{yx}v_{r4} + \tilde{J}_{zx}w_{r4}}_{2R} - \tilde{J}_{yy}2e - \tilde{J}_{zz}2h = 2R - \tilde{J}_{yy}2e + \tilde{J}_{zz}2\frac{\tilde{J}_{zx}}{\tilde{J}_{zy}}g$$

$$0 = C_6 = \underbrace{\tilde{J}_{xy}u_{s4} + \tilde{J}_{yy}v_{s4} + \tilde{J}_{zy}w_{s4}}_{2S} - \tilde{J}_{xx}2b - \tilde{J}_{zz}2g = 2S + \tilde{J}_{xx}2\frac{\tilde{J}_{xy}}{\tilde{J}_{xz}}a - \tilde{J}_{zz}2g \qquad (15.10)$$

This is much more manageable:

$$\begin{pmatrix} \tilde{J}_{xx} & -\tilde{J}_{yy}\tilde{J}_{yz}/\tilde{J}_{yx} & 0 \\ 0 & \tilde{J}_{yy} & -\tilde{J}_{zz}\tilde{J}_{zx}/\tilde{J}_{zy} \\ -\tilde{J}_{xx}\tilde{J}_{xy}/\tilde{J}_{xz} & 0 & \tilde{J}_{zz} \end{pmatrix} \cdot \begin{pmatrix} a \\ e \\ g \end{pmatrix} = \begin{pmatrix} T \\ R \\ S \end{pmatrix}$$

or,

$$\begin{pmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} & A_{23} \\ A_{31} & 0 & A_{33} \end{pmatrix} \cdot \begin{pmatrix} a \\ e \\ g \end{pmatrix} = \begin{pmatrix} T \\ R \\ S \end{pmatrix}$$

The solution is not super-elegant, so I stop here and we might solve the 3x3 system on the fly.

Could there be a case where some off-diagonal $\tilde{J}$ terms are zero and some are not?

Summary

$$
\begin{aligned}
u^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)u_i + (as + bt + c)(1-r)(1+r) \\
v^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)v_i + (dr + et + f)(1-s)(1+s) \\
w^h(r,s,t) &= \sum_i \mathcal{N}_i(r,s,t)w_i + (gr + hs + i)(1-t)(1+t) \\
a &= \ldots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15.11) \\
b &= -\frac{\tilde{J}_{xy}}{\tilde{J}_{xz}}a \\
c &= \frac{1}{2\tilde{J}_{xx}}(\tilde{J}_{xy}u_{s2} + \tilde{J}_{xz}u_{t2} + \tilde{J}_{yy}v_{s2} + \tilde{J}_{yz}v_{t2} + \tilde{J}_{zy}w_{s2} + \tilde{J}_{zz}w_{t2}) \\
d &= -\frac{\tilde{J}_{yz}}{\tilde{J}_{yx}}e \\
e &= \ldots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15.12) \\
f &= \frac{1}{2\tilde{J}_{yy}}(\tilde{J}_{xx}u_{r2} + \tilde{J}_{xz}u_{t3} + \tilde{J}_{yx}v_{r2} + \tilde{J}_{yz}v_{t3} + \tilde{J}_{zx}w_{r2} + \tilde{J}_{zz}w_{t3}) \\
g &= \ldots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15.13) \\
h &= -\frac{\tilde{J}_{zx}}{\tilde{J}_{zy}}g \\
i &= \frac{1}{2\tilde{J}_{zz}}(\tilde{J}_{xx}u_{r3} + \tilde{J}_{xy}u_{s3} + \tilde{J}_{yx}v_{r3} + \tilde{J}_{yy}v_{s3} + \tilde{J}_{zx}w_{r3} + \tilde{J}_{zy}w_{s3})
\end{aligned}
$$

Case of a regular grid made of cuboids In the case of a regular grid with nodes aligned with the x, y, z axis, the 6×6 system above is indefinite as $\tilde{J}_{xy} = \tilde{J}_{yx} = \tilde{J}_{xz} = \ldots = 0$. Let us then rewrite the C equations again in this specific case:

$$
\begin{aligned}
C_0 &= \tilde{J}_{xx}u_{r1} + \tilde{J}_{yy}v_{s1} + \tilde{J}_{zz}w_{t1} \\
C_1 &= \tilde{J}_{yy}v_{s2} + \tilde{J}_{zz}w_{t2} - \tilde{J}_{xx}2c \\
C_2 &= \tilde{J}_{xx}u_{r2} + \tilde{J}_{zz}w_{t3} - \tilde{J}_{yy}2f \\
C_3 &= \tilde{J}_{xx}u_{r3} + \tilde{J}_{yy}v_{s3} - \tilde{J}_{zz}2i \\
C_4 &= \tilde{J}_{zz}w_{t4} - \tilde{J}_{xx}2a - \tilde{J}_{yy}2d \\
C_5 &= \tilde{J}_{xx}u_{r4} - \tilde{J}_{yy}2e - \tilde{J}_{zz}2h \\
C_6 &= \tilde{J}_{yy}v_{s4} - \tilde{J}_{xx}2b - \tilde{J}_{zz}2g \\
C_7 &= 0 \\
C_8 &= 0 \\
C_9 &= 0
\end{aligned}
$$

We see that the condition $C_7 = C_8 = C_9 = 0$ are automatically satisfied. The c, f, i coefficients are obtained as in the general case above. We are left with the equations for C_4, C_5, C_6 (we leave the C_0 equation alone - note that it does not contain any coefficient $a, b, c\ldots$ anymore anyways).

Also, elements are cuboids of size $h_x \times h_y \times h_z$, so that their Jacobian matrix is

$$
\boldsymbol{J} = \begin{pmatrix} h_x/2 & 0 & 0 \\ 0 & h_y/2 & 0 \\ 0 & 0 & h_z/2 \end{pmatrix}
$$

and its inverse:

$$\tilde{J} = \begin{pmatrix} 2/h_x & 0 & 0 \\ 0 & 2/h_y & 0 \\ 0 & 0 & 2/h_z \end{pmatrix}$$

Then the C_4, C_5, C_6 equations become

$$0 = C_4 = \frac{2}{h_z} w_{t4} - \frac{2}{h_x} 2a - \frac{2}{h_y} 2d$$

$$0 = C_5 = \frac{2}{h_x} u_{r4} - \frac{2}{h_y} 2e - \frac{2}{h_z} 2h$$

$$0 = C_6 = \frac{2}{h_y} v_{s4} - \frac{2}{h_x} 2b - \frac{2}{h_z} 2g \qquad (15.14)$$

This is problematic since we are left with 6 unknowns and 3 equations So we should probably go back to the original definition of

$$u^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) u_i + (as + bt + c)(1 - r)(1 + r)$$

$$v^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) v_i + (dr + et + f)(1 - s)(1 + s)$$

$$w^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) w_i + (gr + hs + i)(1 - t)(1 + t)$$

and simply choose 3 of the 6 coefficients a, b, d, e, g, h to be zero ? May be better, as proposed earlier: take $\alpha_r = a = b$, $\alpha_s = d = e$ and $\alpha_t = g = h$? Then, keeping only $\alpha_r, \alpha_s, \alpha_t$:

$$u^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) u_i + (\alpha_r(s + t) + c)(1 - r)(1 + r)$$

$$v^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) v_i + (\alpha_s(r + t) + f)(1 - s)(1 + s)$$

$$w^h(r, s, t) = \sum_i \mathcal{N}_i(r, s, t) w_i + (\alpha_t(r + s) + i)(1 - t)(1 + t)$$

The C_4, C_5, C_6 equations become

$$0 = C_4 = \frac{2}{h_z} w_{t4} - \frac{2}{h_x} 2\alpha_r - \frac{2}{h_y} 2\alpha_s$$

$$0 = C_5 = \frac{2}{h_x} u_{r4} - \frac{2}{h_y} 2\alpha_s - \frac{2}{h_z} 2\alpha_t$$

$$0 = C_6 = \frac{2}{h_y} v_{s4} - \frac{2}{h_x} 2\alpha_r - \frac{2}{h_z} 2\alpha_t$$

and we have 3 equations and 3 unknowns:

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 2h_y/h_x & 0 & 2h_y/h_z \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ v_{s4} \end{pmatrix}$$

multiply last line by h_z/h_y:

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ h_z/h_y \cdot 2h_y/h_x & 0 & h_z/h_y \cdot 2h_y/h_z \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ h_z/h_y \cdot v_{s4} \end{pmatrix}$$

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 2h_z/h_x & 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ h_z/h_y \cdot v_{s4} \end{pmatrix}$$

subtract line 3 from line 1 and put in in line 3:

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 0 & -2h_z/h_y & 2 \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ h_z/h_y \cdot v_{s4} - w_{t4} \end{pmatrix}$$

now multiply 3rd line by h_x/h_z

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 0 & h_x/h_z \cdot -2h_z/h_y & h_x/h_z 2 \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ h_x/h_z(h_z/h_y \cdot v_{s4} - w_{t4}) \end{pmatrix}$$

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 0 & -2h_x/h_y & 2h_x/h_z \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ h_x/h_y \cdot v_{s4} - h_x/h_z w_{t4} \end{pmatrix}$$

Add line 2 to line 3:

$$\begin{pmatrix} 2h_z/h_x & 2h_z/h_y & 0 \\ 0 & 2h_x/h_y & 2h_x/h_z \\ 0 & 0 & 4h_x/h_z \end{pmatrix} \cdot \begin{pmatrix} \alpha_r \\ \alpha_s \\ \alpha_t \end{pmatrix} = \begin{pmatrix} w_{t4} \\ u_{r4} \\ u_{r4} + h_x/h_y \cdot v_{s4} - h_x/h_z w_{t4} \end{pmatrix}$$

From the third line we obtain:

$$\alpha_t = \frac{1}{4}\frac{h_z}{h_x}\left(u_{r4} + \frac{h_x}{h_y}v_{s4} - \frac{h_x}{h_z}w_{t4}\right) = \frac{1}{4}\left(\frac{h_z}{h_x}u_{r4} + \frac{h_z}{h_y}v_{s4} - w_{t4}\right)$$

Then

$$2\frac{h_x}{h_y}\alpha_s + 2\frac{h_x}{h_z}\alpha_t = u_{r4}$$

$$\begin{aligned}
\alpha_s &= \frac{h_y}{h_x}\left(\frac{1}{2}u_{r4} - \frac{h_x}{h_z}\alpha_t\right) \\
&= \frac{1}{2}\frac{h_y}{h_x}u_{r4} - \frac{h_y}{h_z}\alpha_t \\
&= \frac{1}{2}\frac{h_y}{h_x}u_{r4} - \frac{h_y}{h_z}\frac{1}{4}\left(\frac{h_z}{h_x}u_{r4} + \frac{h_z}{h_y}v_{s4} - w_{t4}\right) \\
&= \frac{1}{2}\frac{h_y}{h_x}u_{r4} - \frac{1}{4}\left(\frac{h_y}{h_x}u_{r4} + v_{s4} - \frac{h_y}{h_z}w_{t4}\right) \\
&= \frac{1}{4}\frac{h_y}{h_x}u_{r4} - \frac{1}{4}v_{s4} + \frac{1}{4}\frac{h_y}{h_z}w_{t4} \\
&= \frac{1}{4}\left(\frac{h_y}{h_x}u_{r4} - v_{s4} + \frac{h_y}{h_z}w_{t4}\right)
\end{aligned}$$

and finally:

$$2\frac{h_z}{h_x}\alpha_r + 2\frac{h_z}{h_y}\alpha_s = w_{t4}$$

$$
\begin{aligned}
\alpha_r &= \frac{h_x}{h_z}\left(\frac{1}{2}w_{t4} - \frac{h_z}{h_y}\alpha_s\right) \\
&= \frac{1}{2}\frac{h_x}{h_z}w_{t4} - \frac{h_x}{h_y}\alpha_s \\
&= \frac{1}{2}\frac{h_x}{h_z}w_{t4} - \frac{h_x}{h_y}\left(\frac{1}{4}\frac{h_y}{h_x}u_{r4} - \frac{1}{4}v_{s4} + \frac{1}{4}\frac{h_y}{h_z}w_{t4}\right) \\
&= \frac{1}{2}\frac{h_x}{h_z}w_{t4} - \left(\frac{1}{4}u_{r4} - \frac{1}{4}\frac{h_x}{h_y}v_{s4} + \frac{1}{4}\frac{h_x}{h_z}w_{t4}\right) \\
&= -\frac{1}{4}u_{r4} + \frac{1}{4}\frac{h_x}{h_y}v_{s4} + \frac{1}{4}\frac{h_x}{h_z}w_{t4} \\
&= \frac{1}{4}\left(-u_{r4} + \frac{h_x}{h_y}v_{s4} + \frac{h_x}{h_z}w_{t4}\right)
\end{aligned}
$$

$$
\begin{aligned}
\beta_r &= \frac{1}{2\tilde{J}_{xx}}(\tilde{J}_{yy}v_{s2} + \tilde{J}_{zz}w_{t2}) \\
&= \frac{h_x}{4}\left(\frac{2}{h_y}v_{s2} + \frac{2}{h_z}w_{t2}\right) \\
&= \frac{1}{2}\left(\frac{h_x}{h_y}v_{s2} + \frac{h_x}{h_z}w_{t2}\right) \\
\beta_s &= \frac{1}{2\tilde{J}_{yy}}(\tilde{J}_{xx}u_{r2} + \tilde{J}_{zz}w_{t3}) \\
&= \frac{h_y}{4}\left(\frac{2}{h_x}u_{r2} + \frac{2}{h_z}w_{t3}\right) \\
&= \frac{1}{2}\left(\frac{h_y}{h_x}u_{r2} + \frac{h_y}{h_z}w_{t3}\right) \\
\beta_t &= \frac{1}{2\tilde{J}_{zz}}(\tilde{J}_{xx}u_{r3} + \tilde{J}_{yy}v_{s3}) \\
&= \frac{h_z}{4}\left(\frac{2}{h_x}u_{r3} + \frac{2}{h_y}v_{s3}\right) \\
&= \frac{1}{2}\left(\frac{h_z}{h_x}u_{r3} + \frac{h_z}{h_y}v_{s3}\right)
\end{aligned}
$$

To recap,

$$u^h(r,s,t) = \sum_i \mathcal{N}_i(r,s,t)u_i + (\alpha_r(s+t) + \beta_r)(1-r)(1+r)$$

$$v^h(r,s,t) = \sum_i \mathcal{N}_i(r,s,t)v_i + (\alpha_s(r+t) + \beta_s)(1-s)(1+s)$$

$$w^h(r,s,t) = \sum_i \mathcal{N}_i(r,s,t)w_i + (\alpha_t(r+s) + \beta_t)(1-t)(1+t)$$

$$\alpha_r = \frac{1}{4}\left(-u_{r4} + \frac{h_x}{h_y}v_{s4} + \frac{h_x}{h_z}w_{t4}\right)$$

$$\alpha_s = \frac{1}{4}\left(\frac{h_y}{h_x}u_{r4} - v_{s4} + \frac{h_y}{h_z}w_{t4}\right)$$

$$\alpha_t = \frac{1}{4}\left(\frac{h_z}{h_x}u_{r4} + \frac{h_z}{h_y}v_{s4} - w_{t4}\right)$$

$$\beta_r = \frac{1}{2}\left(\frac{h_x}{h_y}v_{s2} + \frac{h_x}{h_z}w_{t2}\right)$$

$$\beta_s = \frac{1}{2}\left(\frac{h_y}{h_x}u_{r2} + \frac{h_y}{h_z}w_{t3}\right)$$

$$\beta_t = \frac{1}{2}\left(\frac{h_z}{h_x}u_{r3} + \frac{h_z}{h_y}v_{s3}\right)$$

15.2.7   Comparison with Wang *et al.* (2015) for 3D

The following is taken from the supplementary material of Wang *et al.* (2015):

In Eqs (4) and (5), the correction item for each velocity component is calculated based on the other velocity component of the nodes. We extend this approach into 3D situation by adding a quadratic item of x_i to each velocity component:

$$U_i^L(x_1,x_2,x_3) = (1-x_1)(1-x_2)\left[(1-x_3)U_i^a + x_3U_i^e\right] +$$
$$x_1(1-x_2)\left[(1-x_3)U_i^b + x_3U_i^f\right] +$$
$$(1-x_1)\quad x_2\left[(1-x_3)U_i^c + x_3U_i^g\right] +$$
$$x_1\quad x_2\left[(1-x_3)U_i^d + x_3U_i^h\right]$$

(6)

$$\Delta U_1 = x_1(1-x_1)(C_{10} + x_2C_{12}),$$

(7)

$$\Delta U_2 = x_2(1-x_2)(C_{20} + x_3C_{23}),$$

(8)

$$\Delta U_3 = x_3(1-x_3)(C_{30} + x_1C_{31}).$$

(9)

The coefficients of these item ($C_{10}, C_{12}, C_{20}, C_{23}, C_{30}, C_{31}$) in Eqs (7-9) is to be determined. They should satisfy the following divergence free condition for 3D incompressible flow field:

$$\frac{\partial U_1}{\partial x_1} + \frac{\partial U_2}{\partial x_2} + \frac{\partial U_3}{\partial x_3} = 0 \quad . \quad (10)$$

Thus, we take the first derivatives of U_i with respect to x_i based on Eqs (3, 6,7,8,9):

$$\Delta x_1 \frac{\partial U_1}{\partial x_1} = (1-x_2)*(1-x_3)*[U_1^b - U_1^a] + x_2*(1-x_3)*[U_1^d - U_1^c]$$
$$+(1-x_2)*x_3*(U_1^f - U_1^e) + x_2*x_3*(U_1^h - U_1^g)$$
$$+(1-2x_1)*(C_{10} + C_{12}*x_2)$$

(11)

$$\Delta x_2 \frac{\partial U_2}{\partial x_2} = (1 - x_1) * (1 - x_3) * [U_2^c - U_2^a] + x_1 * (1 - x_3) * [U_2^d - U_2^b]$$
$$+ (1 - x_1) * x_3 * (U_2^g - U_2^e) + x_1 * x_3 * (U_2^h - U_2^f)$$
$$+ (1 - 2x_2) * (C_{20} + C_{23} * x_3) \tag{12}$$

$$\Delta x_3 \frac{\partial U_3}{\partial x_3} = (1 - x_1) * (1 - x_2) * [U_3^e - U_3^a] + x_1 * (1 - x_2) * [U_3^f - U_3^b]$$
$$+ (1 - x_1) * x_2 * (U_3^g - U_3^c) + x_1 * x_2 * (U_3^h - U_3^d)$$
$$+ (1 - 2x_3) * (C_{30} + C_{31} * x_1) \tag{13}$$

Substitute Eqs (11-13) in to Eq (10) and we have an identical equation with the six unknowns (C_{10}, C_{12}, C_{20}, C_{23}, C_{30}, C_{31}). As the result, the following items should have their coefficients to be zeros: 1, x_1, x_2, x_3, $x_1 x_2$, $x_2 x_3$, $x_3 x_1$, which lead to 7 equations:

$1:$ $\frac{1}{\Delta x_1}[U_1^b - U_1^a + C_{10}] + \frac{1}{\Delta x_2}[U_2^c - U_2^a + C_{20}] + \frac{1}{\Delta x_3}[U_3^e - U_3^a + C_{30}] = 0$

$x_1:$ $\frac{1}{\Delta x_1}[-2C_{10}] + \frac{1}{\Delta x_2}[U_2^a - U_2^c + U_2^d - U_2^b] + \frac{1}{\Delta x_3}[U_3^a - U_3^e + U_3^f - U_3^b + C_{31}] = 0$

$x_2:$ $\frac{1}{\Delta x_1}[U_1^a - U_1^b + U_1^d - U_1^c + C_{12}] + \frac{1}{\Delta x_2}[-2C_{20}] + \frac{1}{\Delta x_3}[U_3^a - U_3^e + U_3^g - U_3^c] = 0$

$x_3:$ $\frac{1}{\Delta x_1}[U_1^a - U_1^b + U_1^f - U_1^e] + \frac{1}{\Delta x_2}[U_2^a - U_2^c + U_2^g - U_2^e + C_{23}] + \frac{1}{\Delta x_3}[-2C_{30}] = 0$

$x_1 x_2:$ $\frac{1}{\Delta x_1}[-2C_{12}] + \frac{1}{\Delta x_3}[U_3^e - U_3^a + U_3^b - U_3^f + U_3^c - U_3^g + U_3^h - U_3^d] = 0$

$x_2 x_3:$ $\frac{1}{\Delta x_1}[U_1^b - U_1^a + U_1^c - U_1^d + U_1^e - U_1^f + U_1^h - U_1^g] + \frac{1}{\Delta x_2}[-2C_{23}] = 0$

$x_3 x_1:$ $\frac{1}{\Delta x_2}[U_2^c - U_2^a + U_2^b - U_2^d + U_2^e - U_2^g + U_2^h - U_2^f] + \frac{1}{\Delta x_3}[-2C_{31}] = 0$

From (10), we could also have

$$\frac{(U_1^a + U_1^c + U_1^e + U_1^g - U_1^b - U_1^d - U_1^f - U_1^h)}{\Delta x1}$$
$$+\frac{(U_2^a + U_2^b + U_2^e + U_2^f - U_2^c - U_2^d - U_2^g - U_2^h)}{\Delta x2}$$
$$+\frac{(U_3^a + U_3^b + U_3^c + U_3^d - U_3^e - U_3^f - U_3^g - U_3^h)}{\Delta x3} = 0$$

(14)

which reduces 7 equations into 6 independent equations. Therefore, the coefficients in Eqs (7-9) as the six unknowns are determined as follows :

$C_{12} = \frac{\Delta x_1}{2 \Delta x_3}[U_3^e - U_3^a + U_3^b - U_3^f + U_3^c - U_3^g + U_3^h - U_3^d]$,

$C_{23} = \frac{\Delta x_2}{2 \Delta x_1}[U_1^b - U_1^a + U_1^c - U_1^d + U_1^e - U_1^f + U_1^h - U_1^g]$,

$C_{31} = \frac{\Delta x_3}{2 \Delta x_2}[U_2^c - U_2^a + U_2^b - U_2^d + U_2^e - U_2^g + U_2^h - U_2^f]$,

$C_{10} = \frac{\Delta x_1}{2 \Delta x_2}[U_2^a - U_2^c + U_2^d - U_2^b] + \frac{\Delta x_1}{2 \Delta x_3}[U_3^a - U_3^e + U_3^f - U_3^b + C_{31}]$,

$C_{20} = \frac{\Delta x_2}{2 \Delta x_1}[U_1^a - U_1^b + U_1^d - U_1^c + C_{12}] + \frac{\Delta x_2}{2 \Delta x_3}[U_3^a - U_3^e + U_3^g - U_3^c]$,

$C_{30} = \frac{\Delta x_3}{2 \Delta x_1}[U_1^a - U_1^b + U_1^f - U_1^e] + \frac{\Delta x_3}{2 \Delta x_2}[U_2^a - U_2^c + U_2^g - U_2^e + C_{23}]$.

Taken from the supplementary material of Wang *et al.* (2015).

In my opinion, it is quite unbelievable that such a document was accepted for publication (even as supplementary material). There is not much justification for why their equation 7 only contains x_2 and not also x_3, same for the other two equations. Rather surprising is also the fact that although equations 3,6,7,8,9 do not contain any $\Delta x_{\{1,2,3\}}$ term then equations 11,12,13 do feature them. Nevertheless, we must make sense of this mess.

Since the authors state that they "transform the rectangular cells into unit squares" I do away with $\Delta x_1 = \Delta x_2 = \Delta x_3 = 1$ altogether. Also, U_1, U_2, U_3 have become U, V, W.

The polynomial representation of U, V, W on the element including the correction factors is

$$
\begin{aligned}
U &= (1-x_1)(1-x_2)(1-x_3)U^a + (1-x_1)(1-x_2)x_3 U^e \\
&+ x_1(1-x_2)(1-x_3)U^b + x_1(1-x_2)x_3 U^f \\
&+ (1-x_1)x_2(1-x_3)U^c + (1-x_1)x_2 x_3 U^g \\
&+ x_1 x_2(1-x_3)U^d + x_1 x_2 x_3 U^h \\
&+ x_1(1-x_1)(C_{10} + x_2 C_{12}) \quad (15.15) \\
V &= (1-x_1)(1-x_2)(1-x_3)V^a + (1-x_1)(1-x_2)x_3 V^e \\
&+ x_1(1-x_2)(1-x_3)V^b + x_1(1-x_2)x_3 V^f \\
&+ (1-x_1)x_2(1-x_3)V^c + (1-x_1)x_2 x_3 V^g \\
&+ x_1 x_2(1-x_3)V^d + x_1 x_2 x_3 V^h \\
&+ x_2(1-x_2)(C_{20} + x_3 C_{23}) \quad (15.16) \\
W &= (1-x_1)(1-x_2)(1-x_3)W^a + (1-x_1)(1-x_2)x_3 W^e \\
&+ x_1(1-x_2)(1-x_3)W^b + x_1(1-x_2)x_3 W^f \\
&+ (1-x_1)x_2(1-x_3)W^c + (1-x_1)x_2 x_3 W^g \\
&+ x_1 x_2(1-x_3)W^d + x_1 x_2 x_3 W^h \\
&+ x_3(1-x_3)(C_{30} + x_1 C_{31}) \quad (15.17)
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{\partial U}{\partial x_1} &= -(1-x_2)(1-x_3)U^a - (1-x_2)x_3 U^e + (1-x_2)(1-x_3)U^b + (1-x_2)x_3 U^f \\
&+ -x_2(1-x_3)U^c - x_2 x_3 U^g + x_2(1-x_3)U^d + x_2 x_3 U^h \\
&+ (1-2x_1)(C_{10} + x_2 C_{12}) \quad (15.18) \\
\frac{\partial V}{\partial x_2} &= -(1-x_1)(1-x_3)V^a - (1-x_1)x_3 V^e - x_1(1-x_3)V^b - x_1 x_3 V^f \\
&+ (1-x_1)(1-x_3)V^c + (1-x_1)x_3 V^g + x_1(1-x_3)V^d + x_1 x_3 V^h \\
&+ (1-2x_2)(C_{20} + x_3 C_{23}) \quad (15.19) \\
\frac{\partial W}{\partial x_3} &= -(1-x_1)(1-x_2)W^a + (1-x_1)(1-x_2)W^e - x_1(1-x_2)W^b + x_1(1-x_2)W^f \\
&- (1-x_1)x_2 W^c + (1-x_1)x_2 W^g - x_1 x_2 W^d + x_1 x_2 W^h \\
&+ (1-2x_3)(C_{30} + x_1 C_{31}) \quad (15.20)
\end{aligned}
$$

So the velocity divergence can be written

$$
\frac{\partial U}{\partial x_1} + \frac{\partial V}{\partial x_2} + \frac{\partial W}{\partial x_3} = A + Bx_1 + Cx_2 + Dx_3 + Ex_1 x_2 + Fx_2 x_3 + Gx_3 x_1 \quad (15.21)
$$

with

$$
\begin{aligned}
A &= -U^a + U^b + C_{10} - V^a + V^c + C_{20} - W^a + W^e + C_{30} \quad &(15.22) \\
B &= -2C_{10} + V^a - V^b - V^c + V^d + W^a - W^e - W^b + W^f + C_{31} \quad &(15.23) \\
C &= U^a - U^b - U^c + U^d + C_{12} - 2C_{20} + W^a - W^e - W^c + W^g \quad &(15.24) \\
D &= U^a - U^e - U^b + U^f + V^a - V^e - V^c + V^g + C_{23} - 2C_{30} \quad &(15.25) \\
E &= -2C_{12} - W^a + W^e + W^b - W^f + W^c - W^g - W^d + W^h \quad &(15.26) \\
F &= -U^a + U^e + U^b - U^f + U^c - U^g - U^d + U^h - 2C_{23} \quad &(15.27) \\
G &= -V^a + V^e + V^b - V^f + V^c - V^g - V^d + V^h - 2C_{31} \quad &(15.28)
\end{aligned}
$$

A term by term comparison of these equations shows that these are identical to the 7 equations in the supplementary material between Eq. 13 and Eq. 14 (why are these not numbered in the supplementary material?).

Ideally we wish to have all 7 coefficients A to G equal to zero. This leaves us with 7 equations involving 6 unknowns. In other words the system is over constrained and cannot be solved. However the authors seem to interprete this in the exact opposite way by offering yet one more constraint (Eq. 14) which a) is irrelevant b) is not justified (it is indeed related to Eq. 10 but only by taking all C_{ij} coefficients equal to zero and expressed for $x_1 = x_2 = x_3 = 1/2$). Funny enough, that constraint of Eq. 14 is not used further...

From $E = 0, F = 0, G = 0$ we get:

$$C_{12} = \frac{1}{2}(-W^a + W^e + W^b - W^f + W^c - W^g - W^d + W^h) \qquad (15.29)$$

$$C_{23} = \frac{1}{2}(-U^a + U^e + U^b - U^f + U^c - U^g - U^d + U^h) \qquad (15.30)$$

$$C_{31} = \frac{1}{2}(-V^a + V^e + V^b - V^f + V^c - V^g - V^d + V^h) \qquad (15.31)$$

and from $B = 0, C = 0, D = 0$ we get

$$C_{10} = \frac{1}{2}(V^a - V^b - V^c + V^d + W^a - W^e - W^b + W^f + C_{31}) \qquad (15.32)$$

$$C_{20} = \frac{1}{2}(U^a - U^b - U^c + U^d + C_{12} + W^a - W^e - W^c + W^g) \qquad (15.33)$$

$$C_{30} = \frac{1}{2}(U^a - U^e - U^b + U^f + V^a - V^e - V^c + V^g + C_{23}) \qquad (15.34)$$

These 6 expressions are identical to the ones in the paper. However, let us now turn to A:

$$A = -U^a + U^b + C_{10} - V^a + V^c + C_{20} - W^a + W^e + C_{30} \qquad (15.35)$$

$$= -U^a + U^b + \frac{1}{2}(V^a - V^b - V^c + V^d + W^a - W^e - W^b + W^f + C_{31}) \qquad (15.36)$$

$$-V^a + V^c + \frac{1}{2}(U^a - U^b - U^c + U^d + C_{12} + W^a - W^e - W^c + W^g) \qquad (15.37)$$

$$-W^a + W^e + \frac{1}{2}(U^a - U^e - U^b + U^f + V^a - V^e - V^c + V^g + C_{23}) \qquad (15.38)$$

$$= -U^a + U^b + \frac{1}{2}(V^a - V^b - V^c + V^d + W^a - W^e - W^b + W^f) \qquad (15.39)$$

$$+\frac{1}{2}\frac{1}{2}(-V^a + V^e + V^b - V^f + V^c - V^g - V^d + V^h) \qquad (15.40)$$

$$-V^a + V^c + \frac{1}{2}(U^a - U^b - U^c + U^d + W^a - W^e - W^c + W^g) \qquad (15.41)$$

$$+\frac{1}{2}\frac{1}{2}(-W^a + W^e + W^b - W^f + W^c - W^g - W^d + W^h) \qquad (15.42)$$

$$-W^a + W^e + \frac{1}{2}(U^a - U^e - U^b + U^f + V^a - V^e - V^c + V^g+) \qquad (15.43)$$

$$+\frac{1}{2}\frac{1}{2}(-U^a + U^e + U^b - U^f + U^c - U^g - U^d + U^h) \qquad (15.44)$$

$$\neq 0 \qquad (15.45)$$

(easy to prove: for example W^h appears only once)

Once again, we find that the divergence is not identically zero in the element, thereby refuting the statement "Adding these corrections does not improve the order of accuracy of the interpolation (it remains a second-order accurate scheme), but they ensure a divergence-free velocity field over the cell" on page 3 of the article. Their entire paper is based on a false premise.

15.2.8   In 2D with P_1 basis functions - what about triangles?

The reference linear element is:

```
s
|
3
|\
| \
|  \
1-----2 ->r
```
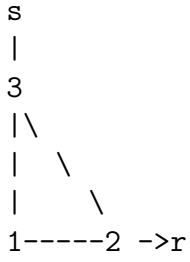
The basis functions are

$$
\begin{array}{rcl}
\mathcal{N}_1(r,s) & = & 1 - r - s \\
\mathcal{N}_2(r,s) & = & r \\
\mathcal{N}_3(r,s) & = & s
\end{array}
\tag{15.46}
$$

and the velocity vector is $\vec{v} = (u, v)$. Its representation inside the element is

$$
\begin{array}{rcl}
u^h(r,s) & = & \displaystyle\sum_i \mathcal{N}_i(r,s) u_i \\[2mm]
v^h(r,s) & = & \displaystyle\sum_i \mathcal{N}_i(r,s) v_i
\end{array}
$$

and the velocity divergence in the element is given by

$$
(\vec{\nabla} \cdot \vec{v})^h = \frac{\partial u^h}{\partial r} + \frac{\partial v^h}{\partial s} = (-u_1 + u_2) + (-v_1 + v_3)
$$

which is evidently not zero everywhere in the element. There is however a fundamental difference with regards to quadrilaterals for which the same quantity still contains r and s terms which opens the door to a correction in order to cancel them. In this case, not so much: this term is exactly the one we could not get rid off for quads!

The following consists of a few misguided attempts at designing a CVI scheme for triangles despite the above observation.

approach 1   As we have seen before the CVI approach consists in adding polynomial terms to the expressions of u^h and v^h. In what follows I assume that the additional terms are of the form (I here use only two basis functions per line, similarly to the quadrilateral counterpart):

$$
u^h(r,s) = \sum_i N_i(r,s) u_i + f(r,s) r (1 - r - s)
\tag{15.47}
$$

$$
v^h(r,s) = \sum_i N_i(r,s) v_i + g(r,s) s (1 - r - s)
\tag{15.48}
$$

Note that we thereby ensure that u is continuous across edges, and so is v.

The velocity divergence requirement is then

$$
\begin{array}{rcl}
0 = \vec{\nabla} \cdot \vec{v}_h & = & -u_1 + u_2 + \partial_r f r (1 - r - s) + f(r,s)(1 - 2r - s) \tag{15.49} \\
& & -v_1 + v_3 + \partial_s g s (1 - r - s) + g(r,s)(1 - 2s - r) \tag{15.50}
\end{array}
$$

- We start simple and postulate $f(r, s) = a$, $g(r, s) = b$, so then

$$
\begin{aligned}
0 = \vec{\nabla} \cdot \vec{v}_h &= -u_1 + u_2 + a(1 - 2r - s) - v_1 + v_3 + b(1 - 2s - r) \qquad (15.51) \\
&= (-u_1 + u_2 - v_1 + v_3 + a + b) + (-2a - b)r + (-a - 2b)s \qquad (15.52)
\end{aligned}
$$

It is impossible to find a and b such that this expression is zero everywhere inside the element.

- We then turn to linear functions and postulate then $f(r, s) = a + br + cs$, $g(r, s) = d + er + fs$, so

$$
\begin{aligned}
0 = \vec{\nabla} \cdot \vec{v}_h &= -u_1 + u_2 + \partial_r f r(1 - r - s) + f(r, s)(1 - 2r - s) \\
&\quad -v_1 + v_3 + \partial_s g s(1 - r - s) + g(r, s)(1 - 2s - r) \\
&= -u_1 + u_2 + br(1 - r - s) + (a + br + cs)(1 - 2r - s) \\
&\quad -v_1 + v_3 + fs(1 - r - s) + (d + er + fs)(1 - 2s - r) \\
&= -u_1 + u_2 - v_1 + v_3 + a + d \\
&\quad +(b - 2a + b - d + e)r \\
&\quad +(f - a + c - 2d + f)s \\
&\quad +(-b - 2b - e)r^2 \\
&\quad +(-f - c - 2f)s^2 \\
&\quad +(-b - f - b - 2c - 2e - f)rs \\
&= -u_1 + u_2 - v_1 + v_3 + a + d \\
&\quad +(2b - 2a - d + e)r \\
&\quad +(2f - a + c - 2d)s \\
&\quad +(-3b - e)r^2 \\
&\quad +(-3f - c)s^2 \\
&\quad +(-2b - 2f - 2c - 2e)rs
\end{aligned}
$$

Immediately $e = -3b$ and $c = -3f$. Inserting these in the last line yields $-2b - 2f - 2c - 2e = -2b - 2f + 6f + 6b = 4b + 4f = 0$, i.e. $b = -f$. Inserting these in the remaining lines:

$$
\begin{aligned}
a + d &= u_1 - u_2 + v_1 - v_3 \\
2b - 2a - d + (-3b) &= 0 \\
2(-b) - a + (3b) - 2d &= 0
\end{aligned}
$$

or,

$$
\begin{aligned}
a + d &= u_1 - u_2 + v_1 - v_3 \\
-2a - b - d &= 0 \\
-a + b - 2d &= 0
\end{aligned}
$$

or,

$$
\begin{pmatrix} 1 & 0 & 1 \\ -2 & -1 & -1 \\ -1 & 1 & -2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ d \end{pmatrix} = \begin{pmatrix} u_1 - u_2 + v_1 - v_3 \\ 0 \\ 0 \end{pmatrix}
$$

Determinant= 3 -2 -1 = 0. Matrix is singular ... !!

- We now try bilinear functions and postulate $f(r, s) = a + br + cs + hrs$, $g(r, s) = d + er + fs + krs$, so then

$$\begin{aligned}
0 = \vec{\nabla} \cdot \vec{v}_h \;=\; & -u_1 + u_2 + \partial_r f r(1 - r - s) + f(r,s)(1 - 2r - s) \\
& -v_1 + v_3 + \partial_s g s(1 - r - s) + g(r,s)(1 - 2s - r) \\
=\; & -u_1 + u_2 + (b + hs)r(1 - r - s) + (a + br + cs + hrs)(1 - 2r - s) \\
& -v_1 + v_3 + (f + kr)s(1 - r - s) + (d + er + fs + krs)(1 - 2s - r) \\
=\; & -u_1 + u_2 - v_1 + v_3 + a + d \\
& +(b - 2a + b - d + e)r \\
& +(f - a + c - 2d + f)s \\
& +(-b - 2b - e)r^2 \\
& +(-f - c - 2f)s^2 \\
& +(-b - f - b - 2c - 2e - f + 2h + 2k)rs \\
& +(-h - k - 2k - h)rs^2 \\
& +(-h - k - 2h - k)r^2 s \\
=\; & -u_1 + u_2 - v_1 + v_3 + a + d \\
& +(b - 2a + b - d + e)r \\
& +(f - a + c - 2d + f)s \\
& +(-3b - e)r^2 \\
& +(-3f - c)s^2 \\
& +(-2b - 2f - 2c - 2e + 2h + 2k)rs \\
& +(-2h - 3k)rs^2 \\
& +(-3h - 2k)r^2 s
\end{aligned} \tag{15.53}$$

Immediately we see that the last 2 lines yield $k = h = 0$ which are the coefficients in front of the new terms (with regards to linear f and g). This is a dead end too.

I *could* keep adding high order terms but I suspect it is a doomed effort and even if it would work, the cost would be prohibitive.

approach 2   This time I include all three basis functions r , s and $1 - r - s$, not just two. Then

$$u^h(r,s) \;=\; \sum_i N_i(r,s)u_i + f(r,s)rs(1 - r - s) \tag{15.54}$$

$$v^h(r,s) \;=\; \sum_i N_i(r,s)v_i + g(r,s)rs(1 - r - s) \tag{15.55}$$

$$\begin{aligned}
0 = \vec{\nabla} \cdot \vec{v}^h \;=\; & -u_1 + u_2 + \partial_r f \; rs(1 - r - s) + f(r,s)s(1 - 2r - s) & \tag{15.56} \\
& -v_1 + v_3 + \partial_s g \; rs(1 - r - s) + g(r,s)r(1 - 2s - r) & \tag{15.57}
\end{aligned}$$

We postulate $f(r,s) = a$, $g(r,s) = b$, so then

$$\begin{aligned}
0 = \vec{\nabla} \cdot \vec{v}^h \;=\; & -u_1 + u_2 + as(1 - 2r - s) - v_1 + v_3 + br(1 - 2s - r) & \tag{15.58} \\
=\; & (-u_1 + u_2 - v_1 + v_3) + ... & \tag{15.59}
\end{aligned}$$

This is also a dead end and this will not change with high order terms in f and g. Because of the presence of all three basis functions in the additional terms we see that no coefficient will enter the parenthesis above and therefore it is doomed.

approach 3 We start from

$$
\begin{pmatrix} \frac{\partial u}{\partial x} \\[2mm] \frac{\partial u}{\partial y} \end{pmatrix} = \tilde{\boldsymbol{J}} \cdot \begin{pmatrix} \frac{\partial u}{\partial r} \\[2mm] \frac{\partial u}{\partial s} \end{pmatrix}
$$

where $\tilde{\boldsymbol{J}}$ in the inverse of the Jacobian matrix. We then postulate again

$$
\begin{aligned}
u(r,s) &= \sum_i \mathcal{N}_i(r,s)u_i + (a_x + b_x r + c_x s + d_x rs + e_x r^2 + f_x s^2) \\
v(r,s) &= \sum_i \mathcal{N}_i(r,s)v_i + (a_y + b_y r + c_y s + d_y rs + e_y r^2 + f_y s^2)
\end{aligned}
$$

In this case,

$$
\frac{\partial u}{\partial r} = \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (b_x + d_x s + 2 e_x r) \tag{15.60}
$$

$$
\frac{\partial u}{\partial s} = \sum_i \frac{\partial \mathcal{N}_i}{\partial s} u_i + (c_x + d_x r + 2 f_x s) \tag{15.61}
$$

$$
\frac{\partial v}{\partial r} = \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (b_y + d_y s + 2 e_y r) \tag{15.62}
$$

$$
\frac{\partial v}{\partial s} = \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (c_y + d_y r + 2 f_y s) \tag{15.63}
$$

We have

$$
\begin{aligned}
\frac{\partial u}{\partial x} &= \tilde{J}_{xx} \frac{\partial u}{\partial r} + \tilde{J}_{xy} \frac{\partial u}{\partial s} \\
&= \tilde{J}_{xx} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i + (b_x + d_x s + 2 e_x r) \right) + \tilde{J}_{xy} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial s} u_i + (c_x + d_x r + 2 f_x s) \right) \\
&= \tilde{J}_{xx} \left(-u_{12} + b_x + d_x s + 2 e_x r \right) \\
&+ \tilde{J}_{xy} \left(-u_{13} + c_x + d_x r + 2 f_x s \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial v}{\partial y} &= \tilde{J}_{yx} \frac{\partial v}{\partial r} + \tilde{J}_{yy} \frac{\partial v}{\partial s} \\
&= \tilde{J}_{yx} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial r} v_i + (b_y + d_y s + 2 e_y r) \right) + \tilde{J}_{yy} \left(\sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i + (c_y + d_y r + 2 f_y s) \right) \\
&= \tilde{J}_{yx} \left(-u_{12} + b_y + d_y s + 2 e_y r \right) \\
&+ \tilde{J}_{yy} \left(-v_{13} + c_y + d_y r + 2 f_y s \right)
\end{aligned}
$$

where $u_{ij} = (u_i - u_j)$ and $v_{ij} = (v_i - v_j)$.

Then

$$
\begin{aligned}
\frac{\partial u^h}{\partial x} + \frac{\partial v^h}{\partial y} &= \tilde{J}_{xx} \left(-u_{12} + b_x + d_x s + 2 e_x r \right) + \tilde{J}_{xy} \left(-u_{13} + c_x + d_x r + 2 f_x s \right) \\
&+ \tilde{J}_{yx} \left(-u_{12} + b_y + d_y s + 2 e_y r \right) + \tilde{J}_{yy} \left(-v_{13} + c_y + d_y r + 2 f_y s \right)
\end{aligned}
$$

We see that yet again velocity components never multiply r nor s so that no space dependent correction can be designed.

15.2.9 In 2D with Q_2 basis functions - Naive approach

```
03===06===02
||   ||   ||
||   ||   ||
07===08===05
||   ||   ||
||   ||   ||
00===04===01
```

The basis functions are given by:

$$
\begin{aligned}
N_0(r, s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \\
N_1(r, s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \\
N_2(r, s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s+1) \\
N_3(r, s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \\
N_4(r, s) &= \frac{1}{2}(1-r^2)s(s-1) \\
N_5(r, s) &= \frac{1}{2}r(r+1)(1-s^2) \\
N_6(r, s) &= \frac{1}{2}(1-r^2)s(s+1) \\
N_7(r, s) &= \frac{1}{2}r(r-1)(1-s^2) \\
N_8(r, s) &= (1-r^2)(1-s^2)
\end{aligned}
$$

and their partial derivatives with respect to the reduced coordinates by

$$\frac{\partial N_0}{\partial r} = \frac{1}{2}(2r - 1)\frac{1}{2}s(s - 1)$$

$$\frac{\partial N_1}{\partial r} = \frac{1}{2}(2r + 1)\frac{1}{2}s(s - 1)$$

$$\frac{\partial N_2}{\partial r} = \frac{1}{2}(2r + 1)\frac{1}{2}s(s + 1)$$

$$\frac{\partial N_3}{\partial r} = \frac{1}{2}(2r - 1)\frac{1}{2}s(s + 1)$$

$$\frac{\partial N_4}{\partial r} = (-2r)\frac{1}{2}s(s - 1)$$

$$\frac{\partial N_5}{\partial r} = \frac{1}{2}(2r + 1)(1 - s^2)$$

$$\frac{\partial N_6}{\partial r} = (-2r)\frac{1}{2}s(s + 1)$$

$$\frac{\partial N_7}{\partial r} = \frac{1}{2}(2r - 1)(1 - s^2)$$

$$\frac{\partial N_8}{\partial r} = (-2r)(1 - s^2)$$

$$\frac{\partial N_0}{\partial s} = \frac{1}{2}r(r - 1)\frac{1}{2}(2s - 1)$$

$$\frac{\partial N_1}{\partial s} = \frac{1}{2}r(r + 1)\frac{1}{2}(2s - 1)$$

$$\frac{\partial N_2}{\partial s} = \frac{1}{2}r(r + 1)\frac{1}{2}(2s + 1)$$

$$\frac{\partial N_3}{\partial s} = \frac{1}{2}r(r - 1)\frac{1}{2}(2s + 1)$$

$$\frac{\partial N_4}{\partial s} = (1 - r^2)\frac{1}{2}(2s - 1)$$

$$\frac{\partial N_5}{\partial s} = \frac{1}{2}r(r + 1)(-2s)$$

$$\frac{\partial N_6}{\partial s} = (1 - r^2)\frac{1}{2}(2s + 1)$$

$$\frac{\partial N_7}{\partial s} = \frac{1}{2}r(r - 1)(-2s)$$

$$\frac{\partial N_8}{\partial s} = (1 - r^2)(-2s)$$

We then have

$$\frac{\partial u^h}{\partial r} = \sum_i \frac{\partial \mathcal{N}_i}{\partial r} u_i$$

$$= \left[\frac{1}{2}(2r-1)\frac{1}{2}s(s-1)\right] u_0 + \left[\frac{1}{2}(2r+1)\frac{1}{2}s(s-1)\right] u_1 + \left[\frac{1}{2}(2r+1)\frac{1}{2}s(s+1)\right] u_2 + \left[\frac{1}{2}(2r-1)\frac{1}{2}s(s+1)\right] u_3$$

$$+ \left[(-2r)\frac{1}{2}s(s-1)\right] u_4 + \left[\frac{1}{2}(2r+1)(1-s^2)\right] u_5 + \left[(-2r)\frac{1}{2}s(s+1)\right] u_6 + \left[\frac{1}{2}(2r-1)(1-s^2)\right] u_7$$

$$+ \left[(-2r)(1-s^2)\right] u_8$$

$$\frac{\partial v^h}{\partial s} = \sum_i \frac{\partial \mathcal{N}_i}{\partial s} v_i$$

$$= \left[\frac{1}{2}r(r-1)\frac{1}{2}(2s-1)\right] v_0 + \left[\frac{1}{2}r(r+1)\frac{1}{2}(2s-1)\right] v_1 + \left[\frac{1}{2}r(r+1)\frac{1}{2}(2s+1)\right] v_2 + \left[\frac{1}{2}r(r-1)\frac{1}{2}(2s+1)\right] v_3$$

$$+ \left[(1-r^2)\frac{1}{2}(2s-1)\right] v_4 + \left[\frac{1}{2}r(r+1)(-2s)\right] v_5 + \left[(1-r^2)\frac{1}{2}(2s+1)\right] v_6 + \left[\frac{1}{2}r(r-1)(-2s)\right] v_7$$

$$+ \left(1-r^2\right)(-2s)\right] v_8$$

or, multiplying each side by 4:

$$4\frac{\partial u^h}{\partial r} = [(2r-1)s(s-1)] u_0 + [(2r+1)s(s-1)] u_1 + [(2r+1)s(s+1)] u_2 + [(2r-1)s(s+1)] u_3$$

$$+ [-4rs(s-1)] u_4 + \left[2(2r+1)(1-s^2)\right] u_5 + [-4rs(s+1)] u_6 + \left[2(2r-1)(1-s^2)\right] u_7$$

$$+ \left[-8r(1-s^2)\right] u_8$$

$$= \left[(2r-1)(s^2-s)\right] u_0 + \left[(2r+1)(s^2-s)\right] u_1 + \left[(2r+1)(s^2+s)\right] u_2 + \left[(2r-1)(s^2+s)\right] u_3$$

$$+ [-4rs(s-1)] u_4 + \left[2(2r+1)(1-s^2)\right] u_5 + [-4rs(s+1)] u_6 + \left[2(2r-1)(1-s^2)\right] u_7$$

$$+ \left[-8r(1-s^2)\right] u_8$$

$$4\frac{\partial v^h}{\partial s} = [r(r-1)(2s-1)] v_0 + [r(r+1)(2s-1)] v_1 + [r(r+1)(2s+1)] v_2 + [r(r-1)(2s+1)] v_3$$

$$+ \left[2(1-r^2)(2s-1)\right] v_4 + [-4rs(r+1)] v_5 + \left[2(1-r^2)(2s+1)\right] v_6 + [-4rs(r-1)] v_7$$

$$+ \left[-8s(1-r^2)\right] v_8$$

$$= \left[(r^2-r)(2s-1)\right] v_0 + \left[(r^2+r)(2s-1)\right] v_1 + \left[(r^2+r)(2s+1)\right] v_2 + \left[(r^2-r)(2s+1)\right] v_3$$

$$+ \left[2(1-r^2)(2s-1)\right] v_4 + [-4rs(r+1)] v_5 + \left[2(1-r^2)(2s+1)\right] v_6 + [-4rs(r-1)] v_7$$

$$+ \left[-8s(1-r^2)\right] v_8$$

We then have

$$4(\vec{\nabla} \cdot \vec{v})^h = 4\frac{\partial u^h}{\partial r} + 4\frac{\partial v^h}{\partial s}$$

$$= (2u_5 - 2u_7 - 2v_4 + 2v_6)\, 1$$

$$+ (4u_5 + 4u_7 - 8u_8 + v_0 - v_1 + v_2 - v_3)\, r$$

$$+ (u_0 - u_1 + u_2 - u_3 + 4v_4 + 4v_6 - 8v_8)\, s$$

$$+ (-2u_0 - 2u_1 + 2u_2 + 2u_3 + 4u_4 - 4u_6 - 2v_0 + 2v_1 + 2v_2 - 2v_3 - 4v_5 + 4v_7)\, rs$$

$$+ (-v_0 - v_1 + v_2 + v_3 + 2v_4 - 2v_6)\, r^2$$

$$+ (-u_0 + u_1 + u_2 - u_3 - 2u_5 + 2u_7)\, s^2$$

$$+ (2v_0 + 2v_1 + 2v_2 + 2v_3 - 4v_4 - 4v_5 - 4v_6 - 4v_7 + 8v_8)\, r^2 s$$

$$+ (2u_0 + 2u_1 + 2u_2 + 2u_3 - 4u_4 - 4u_5 - 4u_6 - 4u_7 + 8u_8)\, rs^2$$

i.e.

$$(\vec{\nabla} \cdot \vec{v})^h \;\; = \;\; C_0 + C_1 r + C_2 s + C_3 rs + C_4 r^2 + C_5 s^2 + C_6 r^2 s + C_7 rs^2 \qquad (15.64)$$

with

$$C_0 \;\; = \;\; \frac{1}{4}(2u_5 - 2u_7 - 2v_4 + 2v_6)$$

$$C_1 \;\; = \;\; \frac{1}{4}(4u_5 + 4u_7 - 8u_8 + v_0 - v_1 + v_2 - v_3)$$

$$C_2 \;\; = \;\; \frac{1}{4}(u_0 - u_1 + u_2 - u_3 + 4v_4 + 4v_6 - 8v_8)$$

$$C_3 \;\; = \;\; \frac{1}{4}(-2u_0 - 2u_1 + 2u_2 + 2u_3 + 4u_4 - 4u_6 - 2v_0 + 2v_1 + 2v_2 - 2v_3 - 4v_5 + 4v_7)$$

$$C_4 \;\; = \;\; \frac{1}{4}(-v_0 - v_1 + v_2 + v_3 + 2v_4 - 2v_6)$$

$$C_5 \;\; = \;\; \frac{1}{4}(-u_0 + u_1 + u_2 - u_3 - 2u_5 + 2u_7)$$

$$C_6 \;\; = \;\; \frac{1}{4}(2v_0 + 2v_1 + 2v_2 + 2v_3 - 4v_4 - 4v_5 - 4v_6 - 4v_7 + 8v_8)$$

$$C_7 \;\; = \;\; \frac{1}{4}(2u_0 + 2u_1 + 2u_2 + 2u_3 - 4u_4 - 4u_5 - 4u_6 - 4u_7 + 8u_8)$$

Looking at C_0, we see that it is effectively $(u_5 - u_7)/2 + (v_6 - v_4)/2$ which is the divergence expressed in the middle of the element using only the mid-edges velocity components (as in a staggered FD grid).

Looking now at C_4 we can write it

$$C_4 = \frac{1}{4}(-(v_0 - 2v_4 + v_1) + (v_3 - 2v_6 + v_2))$$

Since the reference element is of size 2×2, then the distance between nodes 0 and 4, and 4 and 1 respectively is $h = 1$. We then recognise

$$\frac{v_0 - 2v_4 + v_1}{h^2} \sim v_4''$$

and likewise

$$\frac{v_3 - 2v_6 + v_2}{h^2} \sim v_6''$$

Can we recognize more FD stencils?

The divergence inside an element is a polynomial, and as before we then need to design a CVI so that we can get rid of the terms containing the C_{1-7} coefficients (while keeping C_0 as low as possible, although we don't have much control over this).

Because we need that the correction term are zero on the edges ($r = \pm 1$ and $s = \pm 1$), we postulate

$$\delta u(r,s) \;\; = \;\; (1 - r^2) f(r,s)$$
$$\delta v(r,s) \;\; = \;\; (1 - s^2) g(r,s)$$

with

$$f(r,s) \;\; = \;\; \sum_{i=0}^{m} \sum_{j=0}^{n} a_{ij} r^i s^j = a_{00} + a_{10} r + a_{01} s + a_{11} rs + a_{20} r^2 + a_{02} s^2 + a_{12} rs^2 + a_{21} r^2 s + a_{22} r^2 s^2 + \ldots$$

$$g(r,s) \;\; = \;\; \sum_{k=0}^{p} \sum_{l=0}^{q} b_{kl} r^k s^l = b_{00} + b_{10} r + b_{01} s + b_{11} rs + b_{20} r^2 + b_{02} s^2 + b_{12} rs^2 + b_{21} r^2 s + b_{22} r^2 s^2 + \ldots$$

Then the partial derivatives of the velocity corrections are given by:

$$
\begin{aligned}
\frac{\partial}{\partial r}\delta u(r,s) &= -2rf(r,s) + (1-r^2)\frac{\partial f}{\partial r}\\
&= -2r(a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{20}r^2 + a_{02}s^2 + a_{12}rs^2 + a_{21}r^2s + a_{22}r^2s^2 + \ldots)\\
&\quad + (1-r^2)(a_{10} + a_{11}s + 2a_{20}r + a_{12}s^2 + 2a_{21}rs + 2a_{22}rs^2 + \ldots)\\
\frac{\partial}{\partial s}\delta v(r,s) &= -2sg(r,s) + (1-s^2)\frac{\partial g}{\partial s}\\
&= -2s(b_{00} + b_{10}r + b_{01}s + b_{11}rs + b_{20}r^2 + b_{02}s^2 + b_{12}rs^2 + b_{21}r^2s + b_{22}r^2s^2 + \ldots)\\
&\quad + (1-s^2)(b_{01} + b_{11}r + 2b_{02}s + 2b_{12}rs + b_{21}r^2 + 2b_{22}r^2s + \ldots)
\end{aligned}
$$

We immediately see that a_{12}, a_{22}, a_{20}, a_{21}, b_{02}, b_{12}, b_{21} and b_{22} must be zero, as well as all higher order terms because these $r^\alpha s^\beta$ are not present in (15.64). Then

$$
\begin{aligned}
f(r,s) &= a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{02}s^2\\
g(r,s) &= b_{00} + b_{10}r + b_{01}s + b_{11}rs + b_{20}r^2\\
\delta u(r,s) &= (1-r^2)(a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{02}s^2)\\
\delta v(r,s) &= (1-s^2)(b_{00} + b_{10}r + b_{01}s + b_{11}rs + b_{20}r^2)\\
\frac{\partial}{\partial r}\delta u(r,s) &= -2r(a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{02}s^2) + (1-r^2)(a_{10} + a_{11}s)\\
\frac{\partial}{\partial s}\delta v(r,s) &= -2s(b_{00} + b_{10}r + b_{01}s + b_{11}rs + b_{20}r^2) + (1-s^2)(b_{01} + b_{11}r)
\end{aligned}
$$

And we have 10 a_{ij} and b_{kl} coefficients to determine. Let us write the corrected velocity divergence:

$$
\begin{aligned}
(\vec{\nabla}\cdot\vec{v})^h_{CVI} &= (\vec{\nabla}\cdot\vec{v})^h + \frac{\partial}{\partial r}\delta u(r,s) + \frac{\partial}{\partial s}\delta v(r,s)\\
&= C_0 + C_1r + C_2s + C_3rs + C_4r^2 + C_5s^2 + C_6r^2s + C_7rs^2\\
&\quad -2r(a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{02}s^2) + (1-r^2)(a_{10} + a_{11}s)\\
&\quad -2s(b_{00} + b_{10}r + b_{01}s + b_{11}rs + b_{20}r^2) + (1-s^2)(b_{01} + b_{11}r)
\end{aligned}
$$

If we want to cancel all first and second-order polynomial terms we need to have

$$
\begin{aligned}
C_0 + a_{10} + b_{01} &= 0\\
C_1 - 2a_{00} + b_{11} &= 0\\
C_2 + a_{11} - 2b_{00} &= 0\\
C_3 - 2a_{01} - 2b_{10} &= 0\\
C_4 - 3a_{10} &= 0 \qquad\qquad (15.65)\\
C_5 - 3b_{01} &= 0 \qquad\qquad (15.66)\\
C_6 - 3a_{11} - 2b_{20} &= 0\\
C_7 - 2a_{02} - 3b_{11} &= 0
\end{aligned}
$$

In total there are 10 coefficients and 8 only equations. Interestingly, we see that this time around we also do not really stand a chance to actually have $C_0 + a_{10} + b_{01} = 0$ because a_{10} and b_{01} are actually given by (15.65) and (15.66):

$$
\begin{aligned}
a_{10} &= C_4/3\\
b_{01} &= C_5/3
\end{aligned}
$$

654

I am then left with

$$C_1 - 2a_{00} + b_{11} = 0 \tag{15.67}$$
$$C_2 + a_{11} - 2b_{00} = 0 \tag{15.68}$$
$$C_3 - 2a_{01} - 2b_{10} = 0 \tag{15.69}$$
$$C_6 - 3a_{11} - 2b_{20} = 0 \tag{15.70}$$
$$C_7 - 2a_{02} - 3b_{11} = 0 \tag{15.71}$$

I now have 8 unknowns and 5 equations. Since the system is overconstrained, we could further zero b_{20} and a_{02} (thereby removing quadratic terms altogether from f and g). Then (15.70) and (15.71) give

$$a_{11} = C_6/3$$
$$b_{11} = C_7/3$$

and then (15.67) and (15.68) yield

$$a_{00} = \frac{1}{2}(C_1 + b_{11}) = \frac{1}{2}(C_1 + C_7/3)$$
$$b_{00} = \frac{1}{2}(C_2 + a_{11}) = \frac{1}{2}(C_2 + C_6/3)$$

Finally, we are left with (15.69) and we assume for simplicity $a_{01} = b_{10}$ so

$$a_{01} = b_{10} = C_3/4$$

It must be noted that this is only *one* possible approach.

In the end, chosing the a_{ij}'s and b_{kl}'s coefficients as obtained above will yield

$$
\begin{aligned}
(\vec{\nabla} \cdot \vec{v})^h_{CVI} &= C_0 + a_{10} + b_{01} \\
&= C_0 + \frac{C_4}{3} + \frac{C_5}{3} \\
&= \frac{1}{4}(2u_5 - 2u_7 - 2v_4 + 2v_6) + \frac{1}{3}\frac{1}{4}(-v_0 - v_1 + v_2 + v_3 + 2v_4 - 2v_6) + \frac{1}{3}\frac{1}{4}(-u_0 + u_1 + u_2 - u_3 - \\
&= \frac{1}{12}(6u_5 - 6u_7 - 6v_4 + 6v_6 - v_0 - v_1 + v_2 + v_3 + 2v_4 - 2v_6 - u_0 + u_1 + u_2 - u_3 - 2u_5 + 2u_7) \\
&= \frac{1}{12}(-(u_0 + u_3 + u_1 + u_2 - u_3 + 4u_5 - 4u_7 - v_0 - v_1 + v_2 + v_3 - 4v_4 + 4v_6)
\end{aligned}
$$

Finish? What can we say there? what do we recognise?

Finally:

$$
\begin{aligned}
\delta u(r, s) &= (1 - r^2)(a_{00} + a_{10}r + a_{01}s + a_{11}rs + a_{02}s^2) \\
&= (1 - r^2)\left(\frac{1}{2}(C_1 + \frac{C_7}{3}) + \frac{C_4}{3}r + \frac{C_3}{4}s + \frac{C_6}{3}rs\right) \\
&= \frac{1}{12}(1 - r^2)(6C_1 + 2C_7 + 4C_4r + 3C_3s + 4C_6rs) \\
\delta v(r, s) &= (1 - s^2)(b_{00} + b_{10}r + b_{01}s + b_{11}rs) \\
&= (1 - s^2)\left(\frac{1}{2}(C_2 + \frac{C_6}{3}) + \frac{C_3}{4}r + \frac{C_5}{3}s + \frac{C_7}{3}rs\right) \\
&= \frac{1}{12}(1 - s^2)(6C_2 + 2C_6 + 3C_3r + 4C_5s + 4C_7rs)
\end{aligned}
$$

Let us verify one more time:

$$12\frac{\partial}{\partial r}\delta u(r,s) = (-2r)(6C_1 + 2C_7 + 4C_4r + 3C_3s + 4C_6rs) + (1-r^2)(4C_4 + 4C_6s)$$

$$12\frac{\partial}{\partial s}\delta v(r,s) = (-2s)(6C_2 + 2C_6 + 3C_3r + 4C_5s + 4C_7rs) + (1-s^2)(4C_5 + 4C_7r)$$

so that

$$\frac{\partial}{\partial r}\delta u(r,s) + \frac{\partial}{\partial s}\delta v(r,s) = \frac{1}{12}\left(-12C_1r - 4C_7r - 8C_4r^2 - 6C_3rs - 8C_6r^2s + 4C_4 + 4C_6s - 4C_4r^2 - 4C_6r^2s\right)$$

$$\frac{1}{12}\left(-12C_2s - 4C_6s - 6C_3rs - 8C_5s^2 - 8C_7rs^2 + 4C_5 + 4C_7r - 4C_5s^2 - 4C_7rs^2\right)$$

$$= \frac{1}{12}\left(4C_4 + 4C_5 - 12C_1r - 12C_2s - 12C_3rs - 12C_4r^2 - 12C_5s^2 - 12C_6r^2s - 12C_7\right.$$

$$= \frac{1}{3}C_4 + \frac{1}{3}C_5 - C_1r - C_2s - C_3rs - C_4r^2 - C_5s^2 - C_6r^2s - C_7rs^2$$

it adds up!

Recap:

$$
\begin{aligned}
\delta u(r,s) &= \frac{1}{12}(1 - r^2)(6C_1 + 2C_7 + 4C_4 r + 3C_3 s + 4C_6 rs) \\
\delta v(r,s) &= \frac{1}{12}(1 - s^2)(6C_2 + 2C_6 + 3C_3 r + 4C_5 s + 4C_7 rs) \\
C_0 &= \frac{1}{4}(2u_5 - 2u_7 - 2v_4 + 2v_6) \\
C_1 &= \frac{1}{4}(4u_5 + 4u_7 - 8u_8 + v_0 - v_1 + v_2 - v_3) \\
C_2 &= \frac{1}{4}(u_0 - u_1 + u_2 - u_3 + 4v_4 + 4v_6 - 8v_8) \\
C_3 &= \frac{1}{4}(-2u_0 - 2u_1 + 2u_2 + 2u_3 + 4u_4 - 4u_6 - 2v_0 + 2v_1 + 2v_2 - 2v_3 - 4v_5 + 4v_7) \\
C_4 &= \frac{1}{4}(-v_0 - v_1 + v_2 + v_3 + 2v_4 - 2v_6) \\
C_5 &= \frac{1}{4}(-u_0 + u_1 + u_2 - u_3 - 2u_5 + 2u_7) \\
C_6 &= \frac{1}{4}(2v_0 + 2v_1 + 2v_2 + 2v_3 - 4v_4 - 4v_5 - 4v_6 - 4v_7 + 8v_8) \\
C_7 &= \frac{1}{4}(2u_0 + 2u_1 + 2u_2 + 2u_3 - 4u_4 - 4u_5 - 4u_6 - 4u_7 + 8u_8)
\end{aligned}
$$

or

$$
\begin{aligned}
\delta u(r,s) &= (1 - r^2)(a_{00} + a_{10} r + a_{01} s + a_{11} rs) \\
\delta v(r,s) &= (1 - s^2)(b_{00} + b_{10} r + b_{01} s + b_{11} rs) \\
a_{00} &= \frac{1}{2}(C_1 + C_7/3) \\
a_{01} &= C_3/4 \\
a_{10} &= C_4/3 \\
a_{11} &= C_6/3 \\
b_{00} &= \frac{1}{2}(C_2 + C_6/3) \\
b_{01} &= C_5/3 \\
b_{10} &= C_3/4 \\
b_{11} &= C_7/3
\end{aligned}
$$

15.3 The level set function technique

This method was developed in the 80's by Stanley Osher and James Sethian [855].

The Level-set Method (LSM), as it is commonly used in Computational Fluid Dynamics – and especially in Computational Geodynamics – represents a close curve Γ (say, in our case, the interface between two fluids or layers) by means of a function ϕ (called the level-set function, or LSF). Γ is then the zero level-set of ϕ:

$$
\Gamma = \{(x,y) \mid \phi(x,y) = 0\} \tag{15.72}
$$

The convention is that $\phi > 0$ inside the region delimited by Γ and $\phi < 0$ outside. The function value indicates on which side of the interface a point is located (negative or positive) and this is used to identify materials.

Furthermore, if the curve Γ moves with a velocity $\vec{\text{v}}$, then it satisfies the following equation:

$$\frac{\partial \phi}{\partial t} + \vec{\text{v}} \cdot \vec{\nabla}\phi = 0 \tag{15.73}$$

The level set function is generally chosen to be a signed distance function, i.e. $|\vec{\nabla}\phi| = 1$ everywhere and its value is also the distance to the interface. The function value indicates on which side of the interface a point is located (negative or positive) and this is used to identify materials.

As explained in Hillebrand, Thieulot, Geenen, van den Berg, and Spakman [604] (2014), the level-set function ϕ is advected with the velocity $\vec{\text{v}}$ which is obtained by solving the Stokes equations. This velocity does not guarantee that after an advection step the signed distance quality of the LSF is preserved. The LSF then needs to be corrected, which is also called reinitialisation. Finally, solving the advection equation must be done in an accurate manner both in time and space, so that so-called ENO (essentially non-oscillatory) schemes are often employed for the space derivative [1021, 1177].

The level set method has not often been used in the geodynamics community with some notable exceptions. Bourgouin and co-workers use this method combined with Finite Differences to model lava flows [132, 131, 556, 528]. Braun, Thieulot, Fullsack, DeKool, and Huismans [144] (2008) use a so-called particle based level set methodology in their FEM code in conjunction with Adaptive Mesh Refinement. Zlotnik *et al.* coupled the X-FEM method with level set functions to model slab break-off and Rayleigh-Taylor Diapirism [1548]. This same particle level sets are studied by Samuel and Evonuk and applied to geophysical flows [1177]. In Suckale *et al.* (2010) [1302, 1301] the authors investigate simulating buoyancy-driven flow in the presence of large viscosity contrasts. Hale *et al.* (2010) [557] use the LSM in 3D and sudy the dynamics of slab tear faults. An overview of the method and applications can be found in [1020].

Several improvements upon the original LSM have been proposed, such as for instance the conservative level set of Zhao, Bai, Li, and Williams [1516] (2014). The most notable difference between CLS method originally proposed by Olsson *et al.* [1018, 1019] and standard LS method lies in the choice of LS function. Instead of the signed distance function, the CLS methods employ the Heaviside function $H(\phi)$

$$H(\phi) = \begin{cases} 1 & \phi > 0 \\ 1/2 & \phi = 0 \\ 0 & \phi < 0 \end{cases}$$

where ϕ is the signed distance function as in the LSM. In practice, a hyperbolic tangent function is used:

$$H(\phi) = \frac{1}{2}(1 + \tan(\phi/2\epsilon))$$

where ϵ defines the spreading width of H. In the case where there are only two fluids (i.e. a single level set is sufficient), the material properties such as density and viscosity are computed as follows:

$$\rho = \rho_1 + (\rho_2 - \rho_1)H(\phi)$$

$$\eta = \eta_1 + (\eta_2 - \eta_1)H(\phi)$$

🛡 Relevant Literature:

- Review of level-set methods: Frederic Gibou, Ronald Fedkiw, and Stanley Osher. "A review of level-set methods and some recent applications". In: *J. Comp. Phys.* 353 (2018), pp. 82–109. DOI: `10.1016/j.jcp.2017.10.006`

- Interactive 3-D computation of fault surfaces using level sets: B.J. Kadlec, G.A. Dorn, H.M. Tufo, and D.A. Yuen. "Interactive 3-D computation of fault surfaces using level sets". In: *Visual Geosciences* 13.1 (2008), pp. 133–138. DOI: 10.1007/s10069-008-0016-9

- SP Van der Pijl, A Segal, C Vuik, and P Wesseling. "A mass-conserving level-set method for modelling of multi-phase flows". In: *International journal for numerical methods in fluids* 47.4 (2005), pp. 339–361. DOI: 10.1002/d.817

- SP Van der Pijl, A Segal, C Vuik, and P Wesseling. "Computing three-dimensional two-phase flows with a mass-conserving level set method". In: *Computing and Visualization in Science* 11.4-6 (2008), pp. 221–235. DOI: 10.1007/s00791-008-0106-0

- Chohong Min and Frédéric Gibou. "A second order accurate level set method on non-graded adaptive cartesian grids". In: *J. Comp. Phys.* 225.1 (2007), pp. 300–321

- SP Van Der Pijl, A Segal, and C Vuik. "Modelling of three-dimensional bubbly flows with a mass-conserving Level-Set method". In: *Proceedings of the 4th European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS.* 2004

15.4 The field/composition technique

This is the approach taken by the ASPECT developers [776, 593]. Each material i is represented by a compositional field c_i, which takes values between 0 and 1. Each compositional field is then advected with the (prescribed or computed) Stokes velocity [259]:

$$\frac{\partial c_i}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla} c_i = 0 \tag{15.74}$$

The value at a point (Finite element node or quadrature point) is 1 if it is in the domain covered by the material i, and 0 otherwise. In one dimension, each compositional field is a Heavyside function. This approach is somewhat similar to the LSM but the field is essentially discontinuous across the interface, which makes it very difficult to advect. On the plus side, compositional fields need not be reinitialised, as opposed to LSF's.

Accurate numerical advection is a notoriously difficult problem. Unless very specialised techniques are used it often yields undershoot ($c_i < 0$) and overshoot ($c_i > 0$), which ultimately yields mass conservation issues. Also, unless special care is taken, compositional fields tend to become more and more diffuse over time: the SUPG method (Section **??**) and the entropy viscosity method [776, 1150] add small amounts of diffusion to dampen the under- and overshoots. This means that at a given point two or more compositions may have values, which require some form of averaging. If under- and overshoots are present, these averagings can become very problematic and even yield meaningless quantities (e.g. negative viscosities).

One rather old and popular filtering approach is the so-called Lenardic and Kaula (1993) [812] filter:

The filtering algorithm for two-component flow is as follows. An initial step distribution in C is assumed with $C = 0$ and $C = 1$ used to distinguish distinct materials. A high-order upwind solution scheme is applied to equation (2), with prescribed initial conditions, resulting in an uncorrected C field. The field is corrected via the following filtering algorithm:

1. The initial sum of all nodal C values is calculated and is assigned to the variable C_{sum0}.

2. Nodal C values below 0 are set to 0 and the peak value below 0 is assigned to the variable C_{min}.

3. Nodal C values above 1 are set to 1 and the peak value above 1 is assigned to the variable C_{max}.

The basic idea of the filtering algorithm is to ensure that ϕ remains within the bounds $0 \leqslant \phi \leqslant 1$, and to minimize dispersion error. We refer the reader to Lenardic and Kaula (1993) for the detailed explanation and here give the outline of the algorithm for a discrete property field $\phi = \{\phi_i\}$.

Algorithm 10 A property filtering algorithm

(1) Compute the initial sum S_0 of all values of ϕ.
(2) Find the minimal value ϕ_{min} below 0.
(3) Find the maximal value ϕ_{max} above 1.
(4) Set $\phi_i = 0$ if $\phi_i \leqslant |\phi_{min}|$.
(5) Set $\phi_i = 1$ if $\phi_i \geqslant 2 - \phi_{max}$.
(6) Compute the sum S_1 of all values of ϕ.
(7) Compute the number num of $0 < \phi_i < 1$.
(8) Add $dist = (S_1 - S_0)/num$ to all $0 < \phi_i < 1$.

From FENICS book
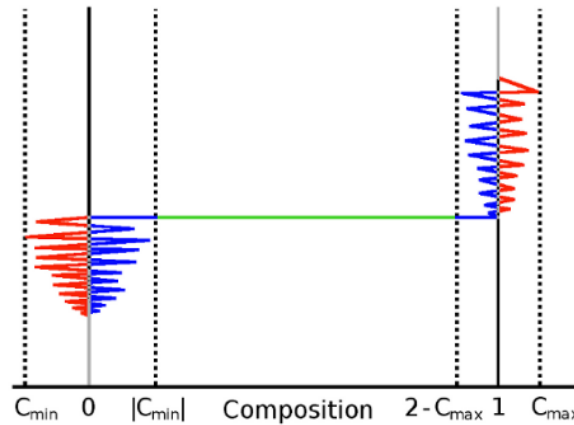
Filtering approach proposed by Lenardic and Kaula (1993). The composition field C is assumed to vary between 0 and 1. Grid points with C-values lower than 0 and greater than 1 are set to 0 and 1, respectively (red). C_{min} and C_{max} are the minimum and maximum spurious values observed. Grid points whose C-value is lower than $|C_{min}|$ or greater than $(2 - C_{max})$ are also set to 0 and 1, respectively (blue). The C-value of all grid points that do not exhibit spurious oscillations (green) is then corrected according to the difference between the original average composition and that computed after the reset-ting of the spurious values. Taken from Plesa *et al.* (2013) [1066].

🗃 Relevant Literature: [1432]
Entropy viscosity method [532]

write about DG approach

15.5 The Volume-of-Fluid method

The Volume-Of-Fluid (VOF) method is a fixed-grid approach based on the one-fluid model and considers that the various immiscible fluids (or 'phases') can be described as a single fluid whose local physical properties, namely density and viscosity, vary in space and time depending on the volume fraction C_i of each phase i [607, 1492].

The volume fraction of each fluid intrinsically obeys $\sum_{i=1}^{n} C_i = 1$ where n is the number of phases. Typically, $C_i = 1$ in grid cells filled only with fluid i, and $0 < C_i < 1$ in grid cells cross–cut by an interface. There are two main classes of VOF methods: methods that try to reconstruct exactly the interface between fluids (e.g. [1085]), which requires significant computational time, and methods that do not, such as in JADIM and OpenFOAM. With no interface reconstruction, the thickness of the interfacial region is defined by $0 < C_i < 1$, and typically occupies two to three grid cells.

♻ Relevant Literature:

Hirt and Nichols [607] "Volume of fluid (VOF) method for the dynamics of free boundaries"
Dutta, Sarkar, and Mandal [378] "Ballooning versus curling of mantle plumes: views from numerical models"
Robey and Puckett [1150] "Implementation of a volume-of-fluid method in a finite element code with applications to thermochemical convection in a density stratified fluid in the earth's mantle"
Louis–Napoleon, Gerbault, Bonometti, Thieulot, Martin, and Vanderhaeghe [857] "3D numerical modeling of crustal polydiapirs with Volume-Of-Fluid methods"
Louis–Napoleon, Bonometti, Gerbault, Martin, and Vanderhaeghe [856] "Models of convection and segregation in heterogeneous partially molten crustal roots with a VOF method–I: flow regimes"

See review of the method in Robey's phd thesis [1149].

15.6 The method of characteristics/Semi-Lagrangian approach

In the syllabus of Becker and Kaus [71] we find in Section 4.8.4 details about the semi-Lagrangian approach. I need to go through it, implement it in a stone and document it here. This approach is also used in Beuchert and Podladchikov [92] (2010).
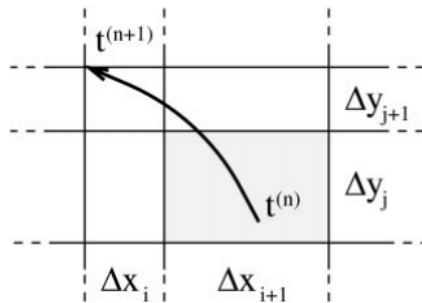
Taken from De Smet, van Den Berg, Vlaar, and Yuen [342] (2000).

In Ilangovan, D'Ascoli, Kohl, and Mohr [652] (2004) we read:

> We approximate T with P_2 elements and apply a splitting approach, where the advective component is resolved with the MMOC [762]. Here virtual particles corresponding to the DoFs of $T_h(t + \delta t)$ are advected back in time along characteristics to obtain their departure points $\boldsymbol{x}_{dept}$ at time t. A Runge-Kutta scheme of order 4 is used to solve the resulting ordinary differential equations (ODEs).

Relevant literature: Nils Kohl, Marcus Mohr, Sebastian Eibl, and Ulrich Rüde. "A massively parallel Eulerian-Lagrangian method for advection-dominated transport in viscous fluids". In: *SIAM J. SCI. COMPUT.* 44.3 (2022), pp. C260–C285. DOI: 10.1137/21M1402510

15.7 The Marker Chain method

In two dimensions, the idea is quite simple: each interface is discretised by means of a number of Lagrangian points (which may or may not vary in time). The points are numbered and connected (think of the connectivity array of a 1D FEM code). In the case of small deformations, and in the absence of in/out-flow boundaries, the method is reasonably trivial to implement, and each couple of point defines a segment (and therefore its normal vector too) which can then be used to answer the question: "at this location, am I above or below this interface" or "am I this domain our outside this domain" (in the case that the interface does not reach any of the boundaries).

This method becomes somewhat impractical when large deformation occurs or, for example, when a domain splits into two (e.g. slab break off). One interface must then become two, and it requires an algorithm capable of detecting the breakup of the surface and capable of rebuilding/patching the new ones so that they can still be used further. Note that in case of large deformation some markers may get further and further apart from each other which makes for a poor representation of the surface. New markers should then be added but the question of when and where must then be addressed.

Also, switching to three dimensions can prove to be very difficult or simply very costly: the generation of the inital marker position is trivial but their connectivity can be complicated to establish at startup: for instance, a Stokes sphere will require a mesh made of triangles which maps exactly the surface of the sphere (see [1344, 957] for methods on how to efficiently produce such meshes). In the case of more complex 3D geometries this may prove nearly impossible to do. So will the problem of splitting a surface into two (or merging two domains).

This method is usually coupled to Eulerian meshes (typically with FDM, but not only). It was used in [1466] in the context of salt domes analysis and later in [266, 263]. It is also used in [1406] but little details are given about the algorithms used to track and update the chain in the presence of such large deformation. It is also used (athough coupled to level set functions) in the DOUAR code[144] (see Section ??). Having worked myself on this code and having had to produce complex initial triangulated surfaces for simulations (see for example [853]) it is easy to understand why later users of this code did implement the marker-in-cell technique. More recently, it is used to track the free surface position in a FDM code [376, 247].

Finally, Christensen [259] makes the following interesting comment: "One might assume that different methods of representing the discontinuity, for example, by a tracer chain [263] or a cloud of tracers, would solve these problems. However, the difficulties arise not only from the way in which material boundaries are represented. Physically, the rate of shear strain parallel to a rheological boundary is discontinuous. Within the finite ele- ment scheme such jump can only be realized at an element boundary. In an Eulerian scheme, where the discontinuity will crosscut the elements, the jump in strain rate must be approx- imated by a continuous variation, and effectively, the rheological properties on both sides of the discontinuity will be averaged in some way within the element."

It is also used in Tan & Gurnis (tagu07) [1319]: " The composition field is computed using the marker chain method [335, 1406]. The marker chain is advected using a fourth-order predictor-corrector scheme. If the distance between two adjacent markers is greater than a predefined threshold, a new marker is inserted in between them. The marker chain defines the material interface. Because of material entrain- ment, the length of the marker chain grows exponentially with time. The computational efficiency of the marker chain method severely deteriorates if there is substantial material entrainment, in which case we halt the computation. For some halted models, the marker chain is trimmed to remove excess entrainment, and the computation restarted in order to proceed further. The trimming of the marker chain introduces error in the composition field, but the magnitude of the error is estimated to be small and does not influence the stability of the chemical layer."

Relevant Literature: Zhong and Hager [1526], Lin and Keken [837], Lin and van Keken [835], Lin and van Keken [836], Kaus [719], Mulyukova [973].

I still have pics from the old days using DOUAR - include

15.8 Hybrid methods

In Braun, Thieulot, Fullsack, DeKool, and Huismans [144] (2008) a level set method is presented which is based on a 3-D set of triangulated points, which makes it a hybrid between tracers and level set functions: in the DOUAR code the interface is then explicitly tracked by means of the tracers while the LSF is computed on the FE nodes. Although very promising in theory, this method proved to be difficult to use in practice since it requires a) a triangulation of the interfaces at $t = 0$ which is not trivial if the geometries are complex (think about a slab in 3D); b) the addition or removal of tracers because of the interface deformation and the patching of the triangulation; c) the calculation of the distance to the interfaces for each FE node based on the triangle normal vectors. This probably explains why the Particle-In-Cell method was later implemented in this code (pers. comm.). Note that another very similar approach is used in [1177].

15.9 Boundary fitted mesh

This method is rather simple to implement and works well for small deformations. It is for instance used by Frehner [443] (see online supplementary material) in which it is stated: "The numerical grid is set up in such a way that the interface between different material phases (two layers in this case) coincides with element boundaries. Hence, each element belongs to a unique material phase and no interpolation is necessary." With such a method, each element is initally attributed a material phase/number and its material properties do not change.

🔰 Relevant Literature: three-dimensional front tracking method using a triangular mesh [1198].

15.10 Runge-Kutta methods

These methods were developed around 1900 by the German mathematicians Carl Runge and Martin Kutta. The RK methods are methods for the numerical integration of ODEs[5]. These methods are well documented in any numerical analysis textbook but also in the textbooks of geodynamics [483, 659]. Any Runge-Kutta method is uniquely identified by its so-called 'Butcher tableau' which contains all necessary coefficients to build the algorithm.

You will find here[6] a complete list of RK methods.

The simplest Runge-Kutta method is the (forward) Euler method. Its tableau is:

$$
\begin{array}{c|c}
0 & \\
\hline
 & 1
\end{array}
$$

The standard second-order RK method method (also called midpoint method) is:

$$
\begin{array}{c|cc}
0 & & \\
1/2 & 1/2 & \\
\hline
 & 0 & 1
\end{array}
$$

Another second-order RK method, called Heun's method[7] is follows:

---

[5] https://en.wikipedia.org/wiki/Runge-Kutta_methods
[6] https://en.wikipedia.org/wiki/List_of_Runge-Kutta_methods
[7] https://en.wikipedia.org/wiki/Heun's_method

$$
\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
& 1/2 & 1/2
\end{array}
$$

A third-order RK method is as follows:

$$
\begin{array}{c|ccc}
0 & & & \\
1/2 & 1/2 & & \\
1 & -1 & 2 & \\
\hline
& 1/6 & 4/6 & 1/6
\end{array}
$$

The RK4 method falls in this framework. Its tableau is:

$$
\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
1/2 & 0 & 1/2 & & \\
1 & 0 & 0 & 1 & \\
\hline
& 1/6 & 1/3 & 1/3 & 1/6
\end{array}
$$

A slight variation of the standard RK4 method is also due to Kutta in 1901 and is called the 3/8-rule. Almost all of the error coefficients are smaller than in the standard method but it requires slightly more FLOPs per time step. Its Butcher tableau is

$$
\begin{array}{c|cccc}
0 & & & & \\
1/3 & 1/3 & & & \\
2/3 & -1/3 & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
& 1/8 & 3/8 & 3/8 & 1/8
\end{array}
$$

We find on Wikipedia[8] a fifth-order method:

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
1/3 & 1/3 & & & & & \\
2/5 & 4/25 & 6/25 & & & & \\
1 & 1/4 & -3 & 15/4 & & & \\
2/3 & 2/27 & 10/9 & -50/81 & 8/81 & & \\
4/5 & 2/25 & 12/25 & 2/15 & 8/75 & 0 & 0 \\
\hline
& 23/192 & 0 & 125/192 & 0 & -27/64 & 125/192
\end{array}
$$

The following method is called the Runge-Kutta-Fehlberg method and is commonly abbreviated RKF45[9]. Its Butcher tableau is as follows:

---

[8] https://en.wikipedia.org/wiki/List_of_Runge-Kutta_methods
[9] https://en.wikipedia.org/wiki/Runge-Kutta-Fehlberg_method

0					
1/4	1/4				
3/8	3/32	9/32			
12/13	1932/2197	-7200/2197	7296/2197		
1	439/216	-8	3680/513	-845/4104	
1/2	-8/27	2	-3544/2565	1859/4104	-11/40
	16/135	0	6656/12825	28561/56430	-9/50	2/55
	25/216	0	1408/2565	2197/4104	-1/5	0

The first row of coefficients at the bottom of the table gives the fifth-order accurate method, and the second row gives the fourth-order accurate method.

The particularity of this method is that from the same Butcher Tableau one can produce a 4th-order approximation $\tilde{A}$ and a 5th-order approximation A:

$$\tilde{A}_{n+1} = A_n + \frac{25}{216}A_1 + \frac{1408}{2565}A_3 + \frac{2197}{4101}A_4 - \frac{1}{5}A_5$$
$$A_{n+1} = A_n + \frac{16}{135}A_1 + \frac{6656}{12825}A_3 + \frac{28561}{56430}A_4 - \frac{9}{50}A_5 + \frac{2}{55}A_6$$

One can define

$$R = \frac{1}{h}|\tilde{A}_{n+1} - A_{n+1}| \qquad \delta = \left(\frac{tol}{R\sqrt{2}}\right)^{1/4}$$

where h is the current step. If $R \leq tol$ keep A as the current step solution and move to the next step with step size $\delta \cdot h$. If $R > tol$ recalculate the current step with step size $\delta \cdot h$.

In the literature we can also find even higher order methods based on the same principle. For example the Butcher tableau of the RKF5 method is as follows[10]:

$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$			
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$		
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$
	$\frac{16}{135}$		$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
	$\frac{25}{216}$		$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	

[10]https://ketch.github.io/numipedia/methods/Fehlberg45.html

c_i	$a_{i,1}$	$a_{i,2}$	$a_{i,3}$	$a_{i,4}$	$a_{i,5}$	$a_{i,6}$	$a_{i,7}$	$a_{i,8}$	$a_{i,9}$	$a_{i,10}$	$a_{i,11}$	$a_{i,12}$
0												
$\frac{2}{27}$	$\frac{2}{27}$											
$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{12}$										
$\frac{1}{6}$	$\frac{1}{24}$	0	$\frac{1}{8}$									
$\frac{5}{12}$	$\frac{5}{12}$	0	$-\frac{25}{16}$	$\frac{25}{16}$								
$\frac{1}{2}$	$\frac{1}{20}$	0	0	$\frac{1}{4}$	$\frac{1}{5}$							
$\frac{5}{6}$	$-\frac{25}{108}$	0	0	$\frac{125}{108}$	$-\frac{65}{27}$	$\frac{125}{54}$						
$\frac{1}{6}$	$\frac{31}{300}$	0	0	0	$\frac{61}{225}$	$-\frac{2}{9}$	$\frac{13}{900}$					
$\frac{2}{3}$	2	0	0	$-\frac{53}{6}$	$\frac{704}{45}$	$-\frac{107}{9}$	$\frac{67}{90}$	3				
$\frac{1}{3}$	$-\frac{91}{108}$	0	0	$\frac{23}{108}$	$-\frac{976}{135}$	$\frac{311}{54}$	$-\frac{19}{60}$	$\frac{17}{6}$	$-\frac{1}{12}$			
1	$\frac{2383}{4100}$	0	0	$\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{301}{82}$	$\frac{2133}{4100}$	$\frac{45}{82}$	$\frac{45}{164}$	$\frac{18}{41}$		
0	$\frac{3}{205}$	0	0	0	0	$-\frac{6}{41}$	$-\frac{3}{205}$	$-\frac{3}{41}$	$\frac{3}{41}$	$\frac{6}{41}$	0	
1	$-\frac{1777}{4100}$	0	0	$\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{289}{82}$	$\frac{2193}{4100}$	$\frac{51}{82}$	$\frac{33}{164}$	$\frac{19}{41}$	0	1
y_1	$\frac{41}{840}$	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	$\frac{41}{840}$	0 $\;\;0$
$\widehat{y}_1$	0	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	0	$\frac{41}{840}$ $\;\;\frac{41}{840}$

7th order Fehlberg method [171].

c_i	a_{ij}							$\hat{b}_i$	b_i
0								$\frac{61}{864}$	$\frac{821}{10800}$
$\frac{1}{10}$	$\frac{1}{10}$							0	0
$\frac{2}{9}$	$\frac{-2}{81}$	$\frac{20}{81}$						$\frac{98415}{321776}$	$\frac{19683}{71825}$
$\frac{3}{7}$	$\frac{615}{1372}$	$\frac{-270}{343}$	$\frac{1053}{1372}$					$\frac{16807}{146016}$	$\frac{175273}{912600}$
$\frac{3}{5}$	$\frac{3243}{5500}$	$\frac{-54}{55}$	$\frac{50949}{71500}$	$\frac{4998}{17875}$				$\frac{1375}{7344}$	$\frac{395}{3672}$
$\frac{4}{5}$	$\frac{-26492}{37125}$	$\frac{72}{55}$	$\frac{2808}{23375}$	$\frac{-24206}{37125}$	$\frac{338}{459}$			$\frac{1375}{5408}$	$\frac{785}{2704}$
1	$\frac{5561}{2376}$	$\frac{-35}{11}$	$\frac{-24117}{31603}$	$\frac{899983}{200772}$	$\frac{-5225}{1836}$	$\frac{3925}{4056}$		$\frac{-37}{1120}$	$\frac{3}{50}$
1	$\frac{465467}{266112}$	$\frac{-2945}{1232}$	$\frac{-5610201}{14158144}$	$\frac{10513573}{3212352}$	$\frac{-424325}{205632}$	$\frac{376225}{454272}$	0	$\frac{1}{10}$	0

c_i	a_{ij}												$\hat{b}_i$	b_i
0													$\frac{14005451}{335480064}$	$\frac{13451932}{455176623}$
$\frac{1}{18}$	$\frac{1}{18}$												0	0
$\frac{1}{12}$	$\frac{1}{48}$	$\frac{1}{16}$											0	0
$\frac{1}{8}$	$\frac{1}{32}$	0	$\frac{3}{32}$										0	0
$\frac{5}{16}$	$\frac{5}{16}$	0	$\frac{-75}{64}$	$\frac{75}{64}$									0	0
$\frac{3}{8}$	$\frac{3}{80}$	0	0	$\frac{3}{16}$	$\frac{3}{20}$								$\frac{-59238493}{1068277825}$	$\frac{-808719846}{976000145}$
$\frac{59}{400}$	$\frac{29443841}{614563906}$	0	0	$\frac{77736538}{692538347}$	$\frac{-28693883}{1125000000}$	$\frac{23124283}{1800000000}$							$\frac{181606767}{758867731}$	$\frac{1757004468}{5645159321}$
$\frac{93}{200}$	$\frac{16016141}{946692911}$	0	0	$\frac{61564180}{158732637}$	$\frac{22789713}{633445777}$	$\frac{545815736}{2771057229}$	$\frac{-180193667}{1043307555}$						$\frac{561292985}{797845732}$	$\frac{656045339}{265891186}$
$\frac{5490023248}{9719169821}$	$\frac{39632708}{573591083}$	0	0	$\frac{-433636366}{683701615}$	$\frac{-421739975}{2616292301}$	$\frac{100302831}{723423059}$	$\frac{790204164}{839813087}$	$\frac{800635310}{3783071287}$					$\frac{-1041891430}{1371343529}$	$\frac{-3867574721}{1518517206}$
$\frac{13}{20}$	$\frac{246121993}{1340847787}$	0	0	$\frac{-37695042795}{15268766246}$	$\frac{-309121744}{1061227803}$	$\frac{-12992083}{490766935}$	$\frac{6005943493}{2108947869}$	$\frac{393006217}{1396673457}$	$\frac{123872331}{1001029789}$				$\frac{760417239}{1151165299}$	$\frac{465885868}{322736535}$
$\frac{1201146811}{1299019798}$	$\frac{-1028468189}{846180014}$	0	0	$\frac{8478235783}{508512852}$	$\frac{1311729495}{1432422823}$	$\frac{-10304129995}{1701304382}$	$\frac{-48777925059}{3047939560}$	$\frac{15336726248}{1032824649}$	$\frac{-45442868181}{3398467696}$	$\frac{3065993473}{597172653}$			$\frac{118820643}{751138087}$	$\frac{53011238}{667516719}$
1	$\frac{185892177}{718116043}$	0	0	$\frac{-3185094517}{667107341}$	$\frac{-477755414}{1098053517}$	$\frac{-703635378}{230739211}$	$\frac{5731566787}{1027545527}$	$\frac{5232866602}{850066563}$	$\frac{-4093664535}{808688257}$	$\frac{3962137247}{1805957418}$	$\frac{65686358}{487910083}$		$\frac{-528747749}{2220607170}$	$\frac{2}{45}$
1	$\frac{403863854}{491063109}$	0	0	$\frac{-5068492393}{434740067}$	$\frac{-411421997}{543043805}$	$\frac{652783627}{914296604}$	$\frac{11173962825}{925320556}$	$\frac{-13158990841}{6184727034}$	$\frac{3936647629}{1978049680}$	$\frac{-160528059}{685178525}$	$\frac{248638103}{1413531060}$	0	$\frac{1}{4}$	0

RK6(5) and RK8(7) Dormand-Prince methods from [1083].

📗 **Relevant Literature:** Dormand and Prince [364] (1980), Fehlberg [414] (1985), Dormand and Prince [365] (1986), Cash and Karp [230] (1990), Hairer, Nørsett, and Wanner [554] (1993), Butcher [209].

Using RK methods to advect particles/markers

In the context of geodynamical modelling, one is usually faced with the following problem: now that I have a velocity field on my FE (or FD) mesh, how can I use it to advect the Lagrangian markers?

Runge-Kutta methods are used to this effect but only their spatial component is used: the velocity solution is not recomputed at the intermediate fractional timesteps, i.e. only the coefficients of the right hand side of the tableaus is used.

- The RK1 method is simple.

$$
\begin{array}{c|c}
0 & \\
\hline
 & 1
\end{array}
$$

    Carry out a loop over markers and

    1. interpolate velocity $\vec{v}_m$ onto each marker m
    2. compute new position as follows: $\vec{r}_m(t + \delta t) = \vec{r}_m(t) + \vec{v}_m \delta t$

- The RK2 method is also simple but requires a bit more work.

$$
\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
 & 1/2 & 1/2
\end{array}
$$

    Carry out a loop over markers and

    1. interpolate velocity $\vec{v}_m$ onto each marker m at position $\vec{r}_m$
    2. compute new intermediate position as follows: $\vec{r}_m^{(1)}(t + \delta t) = \vec{r}_m(t) + \vec{v}_m \delta t/2$
    3. compute velocity $\vec{v}_m^{(1)}$ at position $\vec{r}_m^{(1)}$
    4. compute new position: $\vec{r}_m(t + \delta t) = \vec{r}_m(t) + \vec{v}_m^{(1)} \delta t$

    Note that the intermediate positions could be in a different element of the mesh so extra care must be taken when computing intermediate velocities.

- The RK3 method introduces two intermediate steps.

$$
\begin{array}{c|ccc}
0 & & & \\
1/2 & \frac{1}{2} & & \\
1 & -1 & 2 & \\
\hline
 & \frac{1}{6} & \frac{4}{6} & \frac{1}{6}
\end{array}
$$

    Carry out a loop over markers and

    1. interpolate velocity $\vec{v}_m$ onto each marker m at position $\vec{r}_m$
    2. compute new intermediate position as follows: $\vec{r}_m^{(1)}(t + \delta t) = \vec{r}_m(t) + \frac{1}{2}\vec{v}_m \delta t$
    3. compute velocity $\vec{v}_m^{(1)}$ at position $\vec{r}_m^{(1)}$
    4. compute new intermediate position as follows: $\vec{r}_m^{(2)}(t + \delta t) = \vec{r}_m(t) + (-1\vec{v}_m + 2\vec{v}_m^{(1)})\delta t$
    5. compute velocity $\vec{v}_m^{(2)}$ at position $\vec{r}_m^{(2)}$
    6. compute new position: $\vec{r}_m(t + \delta t) = \vec{r}_m(t) + (\frac{1}{6}\vec{v}_m + \frac{4}{6}\vec{v}_m^{(1)} + \frac{1}{6}\vec{v}_m^{(2)})\delta t$

The following example is borrowed from [871], itself borrowed from Fullsack [452, Section 5.4]. It is a whirl flow [1023], a flow with rotational symmetry in which concentric layers of material rotate around a centre with an angular velocity:

$$
\omega(r) = \omega_0 \frac{r}{r_0} \exp\left(-\frac{r}{r_0}\right)
$$

The box is $[-0.5, 0.5] \times [-0.5, 0.5]$, $r_0 = 0.25$, $\omega_0 = 0.3$ and $\delta t = 1$. 60×60 particles are regularly positioned inside the $[-0.3, 0.3] \times [-0.3, 0.3]$ square. Maierova [871] has carried out this experiment for the above Runge-Kutta methods.
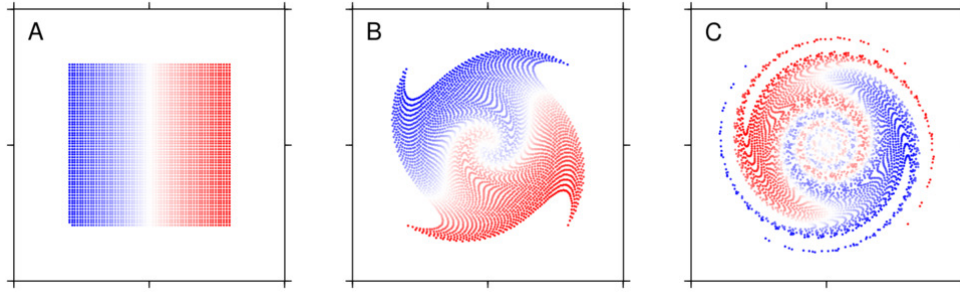
Model domain with particles colored at three different time-steps: (A) t = 0 (initial position of particles), (B) t = 50, and (C) t = 200. The advection is computed using the fourth-order Runge-Kutta scheme. Taken from [871]
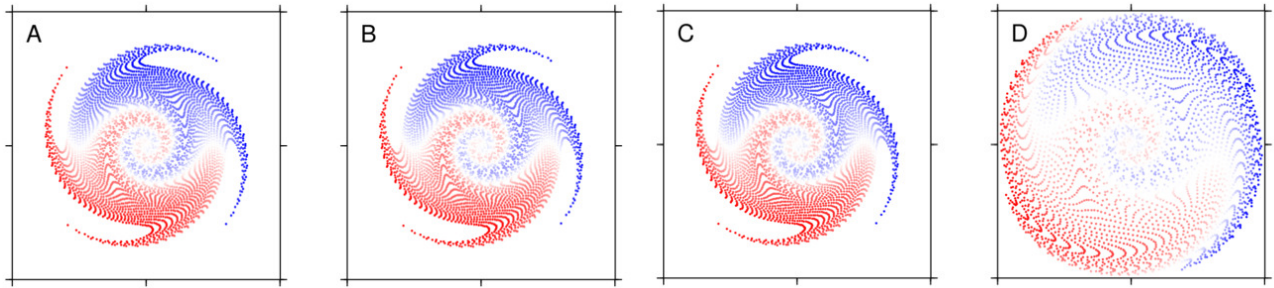
The same plot as above, but for different advection schemes at t = 100. Advection was computed using (A) the fourth-order Runge-Kutta scheme, (B) the mid- point method, (C) Heun's method and (D) the explicit Euler method. Taken from [871]

15.11 Am I in or not? - finding reduced coordinates

It is quite common that at some point one must answer the question: "Given a mesh and its connectivity on the one hand, and the coordinates of a point on the other, how do I accurately and quickly determine in which element the point resides?"

One typical occurence of such a problem is linked to the use of the Particle-In-Cell technique: particles are advected and move through the mesh, and need to be localised at every time step. This question could arise in the context of a benchmark where certain quantities need to be measured at specific locations inside the domain.

Two-dimensional space

We shall first focus on quadrilaterals. There are many kinds of quadrilaterals as shown hereunder:

668
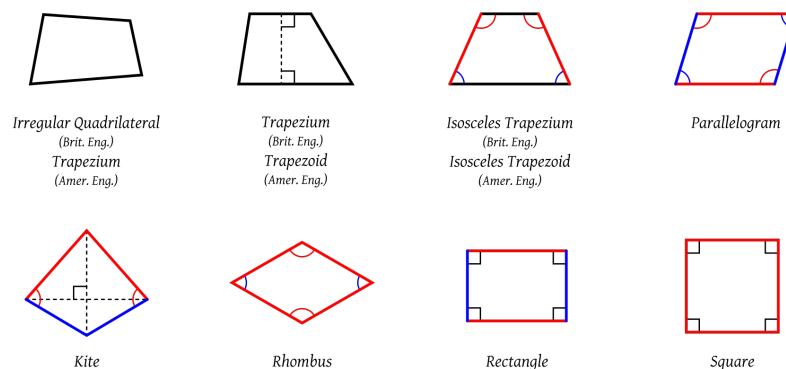
The trivial case of rectangular elements Testing whether the point M is inside the element is trivial. For $x_0 \leq x_M \leq x_2$ and $y_0 \leq y_M \leq y_2$, its reduced coordinates are given by

$$
\begin{aligned}
r_M &= \frac{2}{x_2 - x_0}(x_M - x_0) - 1 = \frac{2}{h_x}(x_M - x_0) - 1 \\
s_M &= \frac{2}{y_2 - y_0}(y_M - y_0) - 1 = \frac{2}{h_y}(y_M - y_0) - 1
\end{aligned}
\tag{15.75}
$$
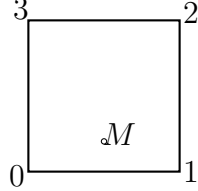
An intermediate case We make the following assumption that the lateral sides of the element are vertical while the bottom and top are not necessarily horizontal:
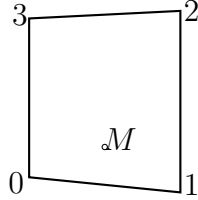
Because the sides are verical then if $x_0 \leq x_M \leq x_2$ then

$$
r_M = \frac{2}{x_2 - x_0}(x_M - x_0) - 1
$$

Then, if M is inside the element then its y coordinate is given by

$$
y_M = \sum_i \mathcal{N}_i(r_M, s_M) y_i
$$

where $\mathcal{N}_i$ are the four Q_1 basis functions associated to the vertices. Assuming we know r_M then we can solve for s_M:

$$
\begin{aligned}
y_M &= \frac{1}{4}(1 - r_M)(1 - s_M)y_0 + \frac{1}{4}(1 + r_M)(1 - s_M)y_1 + \frac{1}{4}(1 + r_M)(1 + s_M)y_2 + \frac{1}{4}(1 - r_M)(1 + s_M)y_3 \\
&= \frac{1}{4}[(1 - r)y_0 + (1 + r)y_1 + (1 + r)y_2 + (1 - r)y_3 + s_M[-(1 - r)y_0 - (1 + r)y_1 + (1 + r)y_2 + (1 - r)y_3]
\end{aligned}
$$

or,

$$
s_M = \frac{4y_M - [(1 - r_M)y_0 + (1 + r_M)y_1 + (1 + r_M)y_2 + (1 - r_M)y_3]}{-(1 - r_M)y_0 - (1 + r_M)y_1 + (1 + r_M)y_2 + (1 - r_M)y_3}
$$

If the obtained value is in $[-1, 1]$ then the point M is in the element. Verification: when $y_1 = y_0$ and

$y_2 = y_3$ then

$$
\begin{aligned}
s_M &= \frac{4y_M - [(1 - r_M)y_0 + (1 + r_M)y_0 + (1 + r_M)y_3 + (1 - r_M)y_3]}{-(1 - r_M)y_0 - (1 + r_M)y_0 + (1 + r_M)y_3 + (1 - r_M)y_3} \\
&= \frac{4y_M - [2y_0 + 2y_3]}{-2y_0 + 2y_3} \\
&= \frac{1}{y_3 - y_0}[2y_M - (y_0 + y_3)] \\
&= \frac{1}{y_3 - y_0}[2y_M - 2y_0 + y_0 - y_3)] \\
&= \frac{2}{y_3 - y_0}(y_M - y_0) - 1 \quad\quad\quad\quad (15.76)
\end{aligned}
$$

which is the expression that corresponds to a rectangular element as seen previously.

A generic quadrilateral   We wish to arrive at a single algorithm which is applicable to all quadrilaterals and we now focus on an irregular quadrilateral (no face is parallel to the axis of the coordinate system).
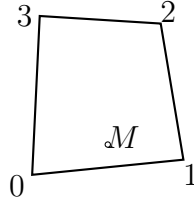
Several rather simple options exist:

- we could subdivide the quadrilateral into two triangles and check whether point M is inside any of them (as it turns out, this problem is rather straightforward for triangles. Simply google it.)

- We could check that point M is always on the left side of segments $0 \to 1$, $1 \to 2$, $2 \to 3$, $3 \to 0$.

- ...

Any of these approaches will work although some might be faster than others. In three-dimensions all will however become cumbersome to implement and might not even work at all. Fortunately, there is an elegant way to answer the question, as detailed in the following subsection, which works both in 2D and 3D.

Three-dimensional space

If point M is inside the quadrilateral, there exist a set of reduced coordinates $r, s, t \in [-1 : 1]^3$ such that

$$
\sum_{i=1}^{4} \mathcal{N}_i(r_M, s_M, t_M)x_i = x_M \quad\quad \sum_{i=1}^{4} \mathcal{N}_i(r_M, s_M, t_M)y_i = y_M \quad\quad \sum_{i=1}^{4} \mathcal{N}_i(r_M, s_M, t_M)z_i = z_M
$$

This can be cast as a system of three equations and three unknowns. Unfortunately, each basis function $\mathcal{N}_i$ contains a term rst (as well as rs, rt, and st) so that it is not a linear system. We must

then use an iterative technique: the algorithm starts with a guess for values r_M, s_M, t_M and improves on their value iteration after iteration. In what follows the subscript M is dropped from r, s, t.

The classical way of solving nonlinear systems of equations is Newton's method. We can rewrite the equations above as $\boldsymbol{F}(r, s, t) = 0$:

$$
\begin{aligned}
\sum_{i=1}^{8} \mathcal{N}_i(r, s, t) x_i - x_M &= 0 \\
\sum_{i=1}^{8} \mathcal{N}_i(r, s, t) y_i - y_M &= 0 \\
\sum_{i=1}^{8} \mathcal{N}_i(r, s, t) z_i - z_M &= 0
\end{aligned}
\tag{15.77}
$$

or,

$$
\begin{aligned}
F_r(r, s, t) &= 0 \\
F_s(r, s, t) &= 0 \\
F_t(r, s, t) &= 0
\end{aligned}
$$

so that we now have to find the zeroes of continuously differentiable functions $\boldsymbol{F} : \mathbb{R} \to \mathbb{R}$. The recursion is simply:

$$
\begin{pmatrix} r_{k+1} \\ s_{k+1} \\ t_{k+1} \end{pmatrix} = \begin{pmatrix} r_k \\ s_k \\ t_k \end{pmatrix} - J_F(r_k, s_k, t_k)^{-1} \begin{pmatrix} F_r(r_k, s_k, t_k) \\ F_s(r_k, s_k, t_k) \\ F_t(r_k, s_k, t_k) \end{pmatrix}
$$

where J the Jacobian matrix:

$$
\begin{aligned}
J_F(r_k, s_k, t_k) &= \begin{pmatrix} \frac{\partial F_r}{\partial r}(r_k, s_k, t_k) & \frac{\partial F_r}{\partial s}(r_k, s_k, t_k) & \frac{\partial F_r}{\partial t}(r_k, s_k, t_k) \\[2mm] \frac{\partial F_s}{\partial r}(r_k, s_k, t_k) & \frac{\partial F_s}{\partial s}(r_k, s_k, t_k) & \frac{\partial F_s}{\partial t}(r_k, s_k, t_k) \\[2mm] \frac{\partial F_t}{\partial r}(r_k, s_k, t_k) & \frac{\partial F_t}{\partial s}(r_k, s_k, t_k) & \frac{\partial F_t}{\partial t}(r_k, s_k, t_k) \end{pmatrix} \\[4mm]
&= \begin{pmatrix} \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial r}(r_k, s_k, t_k) x_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial s}(r_k, s_k, t_k) x_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial t}(r_k, s_k, t_k) x_i \\[2mm] \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial r}(r_k, s_k, t_k) y_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial s}(r_k, s_k, t_k) y_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial t}(r_k, s_k, t_k) y_i \\[2mm] \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial r}(r_k, s_k, t_k) z_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial s}(r_k, s_k, t_k) z_i & \sum_{i=1}^{8} \frac{\partial \mathcal{N}_i}{\partial t}(r_k, s_k, t_k) z_i \end{pmatrix}
\end{aligned}
$$

In practice, we solve the following system:

$$
J_F(r_k, s_k, t_k) \left[\begin{pmatrix} r_{k+1} \\ s_{k+1} \\ t_{k+1} \end{pmatrix} - \begin{pmatrix} r_k \\ s_k \\ t_k \end{pmatrix} \right] = - \begin{pmatrix} F_r(r_k, s_k, t_k) \\ F_s(r_k, s_k, t_k) \\ F_t(r_k, s_k, t_k) \end{pmatrix}
$$

Finally, the algorithm goes as follows:

- set guess values for r, s, t (typically 0)

- loop over k=0,...

- Compute rhs$= -\boldsymbol{F}(r_k, s_k, t_k)$

- Compute matrix $J_F(r_k, s_k, t_k)$

- solve system for (dr_k, ds_k, dt_k)

- update $r_{k+1} = r_k + dr_k$, $s_{k+1} = s_k + ds_k$, $t_{k+1} = t_k + dt_k$

- stop iterations when (dr_k, ds_k, dt_k) is small

- if $r_k, s_k, t_k \in [-1, 1]^3$ then M is inside.

This method converges quickly but involves iterations, and multiple solves of 3×3 systems which, when carried out for each marker and at each time step can prove to be expensive. A simple modification can be added to the above algorithm: iterations should be carried out *only* when the point M is inside of a cuboid of size $[\min_i x_i : \max_i x_i] \times [\min_i y_i : \max_i y_i] \times [\min_i z_i : \max_i z_i]$ where the sums run over the vertices of the element. In 2D this translates as follows: only carry out Newton iterations when M is inside the red rectangle!
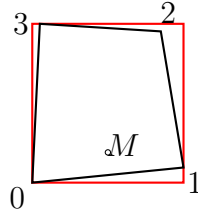
Note that the algorithm above extends to high degree elements such as Q_2 and higher, even with curved sides. As shown in the 2D case if the element is a cuboid or if all its lateral faces are vertical then one can compute the reduced coordinates without using an iterative method.

Three-dimensional space - special case

We assume that the mesh is such that the cross section of all Q_1 elements is a rectangle in the xy-plane.

Let (x, y, z) be a point inside the element. The global coordinates x, y, z are obtained from the reduced coordinates r, s, t via the basis the basis functions:

$$x = \sum_{i=1}^{8} \mathcal{N}_i(r, s, t) x_i \qquad y = \sum_{i=1}^{8} \mathcal{N}_i(r, s, t) y_i \qquad z = \sum_{i=1}^{8} \mathcal{N}_i(r, s, t) z_i \qquad (15.78)$$

Let

$$
\begin{aligned}
\vec{v}_1 &= (+1, +1, +1, +1, +1, +1, +1, +1) \\
\vec{v}_2 &= (-1, +1, +1, -1, -1, +1, +1, -1) \\
\vec{v}_3 &= (-1, -1, +1, +1, -1, -1, +1, +1) \\
\vec{v}_4 &= (-1, -1, -1, -1, +1, +1, +1, +1) \\
\vec{v}_5 &= (+1, -1, +1, -1, +1, -1, +1, -1) \\
\vec{v}_6 &= (+1, -1, -1, +1, -1, +1, +1, -1) \\
\vec{v}_7 &= (+1, +1, -1, -1, -1, -1, +1, +1)
\end{aligned}
$$

and

$$
\begin{aligned}
\vec{x} &= (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \\
\vec{y} &= (y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8) \\
\vec{z} &= (z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8)
\end{aligned}
$$

672

then Eqs. (15.78) can also be written

$$x = \frac{1}{8} (\vec{v}_1 + r\vec{v}_2 + s\vec{v}_3 + t\vec{v}_4 + rs\vec{v}_5 + rt\vec{v}_6 + st\vec{v}_7) \cdot \vec{x}$$

$$y = \frac{1}{8} (\vec{v}_1 + r\vec{v}_2 + s\vec{v}_3 + t\vec{v}_4 + rs\vec{v}_5 + rt\vec{v}_6 + st\vec{v}_7) \cdot \vec{y}$$

$$z = \frac{1}{8} (\vec{v}_1 + r\vec{v}_2 + s\vec{v}_3 + t\vec{v}_4 + rs\vec{v}_5 + rt\vec{v}_6 + st\vec{v}_7) \cdot \vec{z} \qquad (15.79)$$

If the element has a rectangular cross-section $s_x \times s_y$ then

$$\vec{x} = (x_0, x_0 + s_x, x_0 + s_x, x_0, x_0, x_0 + s_x, x_0 + s_x, x_0)$$
$$\vec{y} = (y_0, y_0, y_0 + s_y, y_0 + s_y, y_0, y_0, y_0 + s_y, y_0 + s_y)$$

which yields

$$r = 2\frac{x - x_0}{s_x} - 1$$

$$s = 2\frac{y - y_0}{s_y} - 1$$

Since the local coordinates r and s can be easily computed, one can use Eq. (15.79) to obtain t:

$$t = \frac{8z - (\vec{v}_1 + r\vec{v}_2 + s\vec{v}_3 + rs\vec{v}_5) \cdot \vec{z}}{(\vec{v}_4 + r\vec{v}_6 + s\vec{v}_7) \cdot \vec{z}}$$

Chapter 16

Additional techniques, features, measurements

Solving the Stokes equations and the energy equations is one thing. Doing it in a geodynamical context requires a lot of additional techniques.

16.1 Dealing with a free surface (and mesh deformation)

When carrying out global models, typically mantle convection, the effect of the free surface is often neglected/negligeable: topography ranges from ~ 10km depth to ~ 10km height, which is very small compared to the depth of the mantle (~ 3000km).

However, it has long been regognised that there is a feedback between topography and crust/lithosphere deformation: the surface of the Earth reflects the deeper processes, from orogeny, back-arc basins, rifts, mid-ocean ridges, etc ... (see for instance Braun (2010) [146]).

Remark. *Free surface flows are not unique to Earth sciences, and their modelling has given rise to many studies and textbooks. A typical free-surface flow problem in the CFD literature is the so-called 'dam break' problem [939, 46, 834, 802, 617, 17]. Other occurrences involve sea waves, flow over structures, flow around ships, mould filling, flow with bubbles [834].*

Remark. *Free surface flows have also been extensively studied and even benchmarked in the ice-sheet modelling community, see [XXX].*

What distinguishes geodynamics free surface modelling from its engineering counterpart is (i) the absence of surface tension, (ii) the fact that the fluids under consideration are Stokesian, (iii) their rheology is complex (the elastic and plastic components can be dominant at the surface).

The problem of dealing with a free surface can be deceptively simple at first glance: as mentioned before the amplitude of surface movement is often less than 1% of the domain size. Isostacy-driven movements are easy to deal with since the movement is vertical (and often characteried by long wavelength). However, computational problems quickly arise for example in subduction modelling: the downgoing lithosphere subducts below the overriding plate and the relative convergence of the two is likely to generate a cusp at the trench. The presence of shear bands intersecting the surface accentuates the problem:
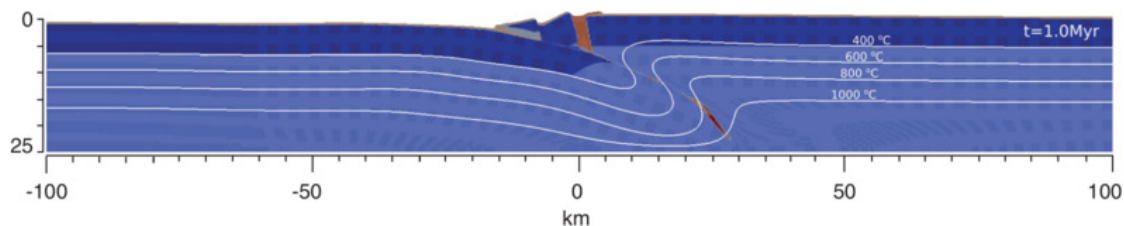
Taken from Maffione *et al.* [869]. Example of free surface deformation above intra-oceanic subduction initiation

Remark. *It is difficult to talk about free surface without including the underlying mesh. What follows should be read alongside Section 16.7.*

The fully Lagrangian approach

In this case the mesh is deformed with the velocity (or displacement) computed on its nodes. It is sometimes called 'body fitting' [304] or 'boundary fitted'. In the case when large deformation occurs (which is rather frequent in geodynamics – think about subduction or rifting processes where materials end up moving 100's or 1000's of km, horizontally and/or vertically) – it leads to highly deformed elements, and in some case even bow-tied:
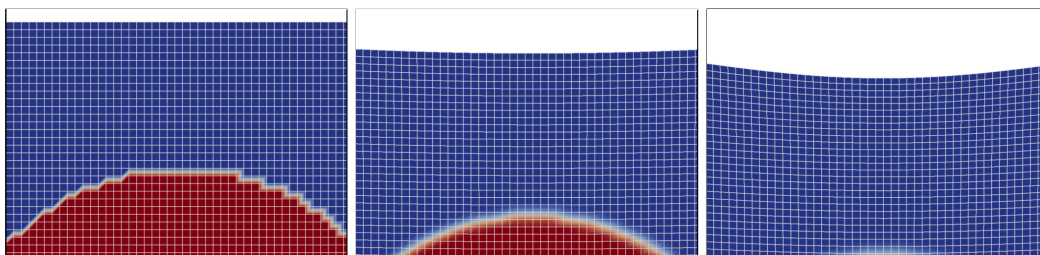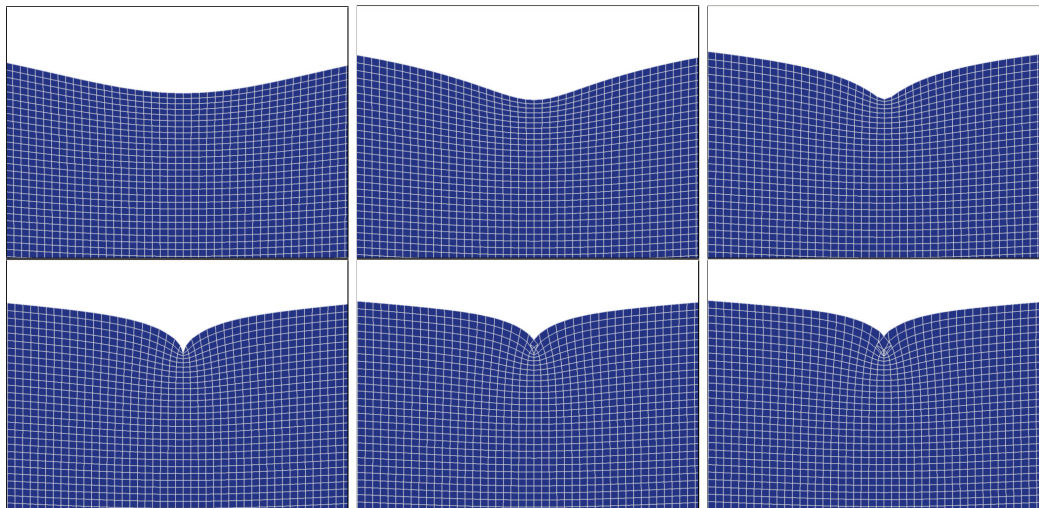
Example of a free surface evolution above a sinking sphere. The isostatic rebound above the sphere generates a cusp which, if no special measure is taken, ultimately leads to a bow-tied element. Once this occurs the simulation stops since the mapping of the bow-tied element to the reference element yields to wrong elemental matrix. Curtesy of M. Fraters

In the mildest cases this does not occur but it has long been established that large mesh deformation yields low accuracy calculations, especially when angles between edges become small or large. One way to overcome this problem is to remesh, i.e. generate a better mesh based on the available information on the deformed one. In 2D this is routinely done, especially when triangular elements are used. In 3D, multiple remeshing are very costly and it is generally avoided. Note also that re-meshing often involves some form of interpolation and therefore some unwanted numerical diffusion. When deformation is reasonably small, fully lagrangian methods work and have been used in geodynamics, see for example Hassani & Chéry (1996) [587], Melosh & Raefsky (1980) [911], or Lavier *et al.* (2000) [799].
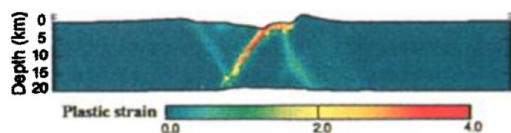
Taken from Lavier *et al.* (2000) [799]. Upper-crustal faulting, note that the bottom and the top surface are deformed.
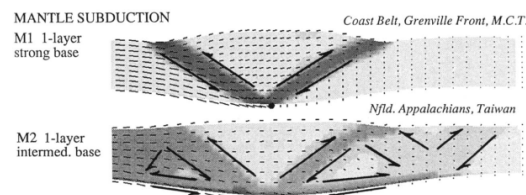
Taken from Beaumon *et al.* (1994) [65]. Strain rate and velocity field for crustal S-point models.
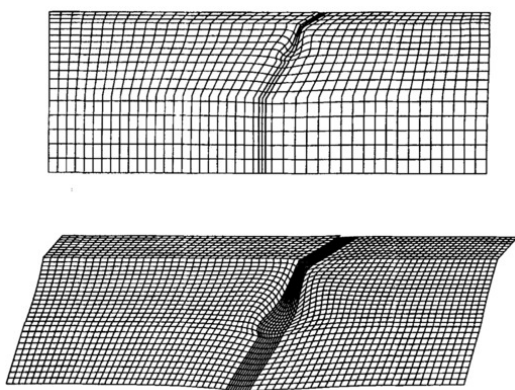
Taken from Gurnis *et al.* (1996) [546]. Subduction model, topographic expression is shown without vertical exaggeration.
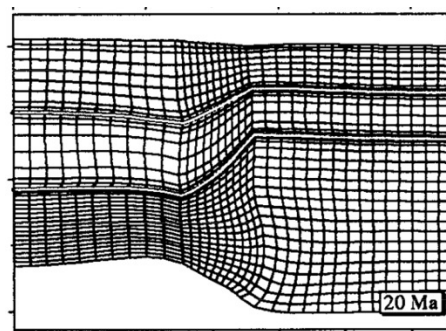
Taken from Govers & Wortel [508]. Asymmetric lithospheric extension.

The Eulerian approach: using sticky air

'Sticky air' is the default option for numerical methods which mesh cannot be deformed (typically the finite difference method). In this case, the air above the crust/sediments is modelled as a zero-

density[1] fluid with very low viscosity (see for instance the early article by Zaleski and Julien [1501]). One problem quickly arises when one realises that the viscosity of the air ($\sim 18.5 \cdot 10^{-6}$ Pa·s[2]) is almost 25-30 orders of magnitude lower than the (effective) viscosity of Earth materials. Real air viscosity cannot therefore be used because of 1) round-off errors, 2) extremely poorly-conditioned matrices. Low viscosities around $10^{16} - 10^{19}$ Pa s are then commonly used as they are still negligible next to those of the (plastic) crust, and the flow of air parallel to Earth materials only generates extremely small shear and normal stress values (thereby approaching the true nature of a free surface). This approach is the one employed in all the papers based on the I2/I3(EL)VIS code (see Appendix **??**) and has been benchmarked in Crameri *et al.* (2012) [304].

This approach has a few advantages:

1. it is simple to implement

2. it is compatible with all the standard numerical methods (FEM, FDM,FVM)

3. it avoids (potentially complicated or costly) remeshing

and quite a few drawbacks:

1. it increases the size of the computational domain, thereby adding more unknowns to the linear system: in Schmeling *et al.* (2008) [1200] the air layer is set to 50 km while the lithospheric domain underneath is 700 km thick;

2. it requires the use of averaging all along the free-surface where very large viscosity contrasts are present. Here is what Poliakov and Podlachikov [1073] say about the sticky air method: "Zaleski & Julien [1501] used a top layer with a very low viscosity and density to represent air or water above the surface. This allows a simple representation of the free surface. However, due to the very high viscosity and density contrast and diffusion between the top layer and the underlying layers, calculations sometimes become unstable and give significant errors."

3. it can yield air entrainment in the mantle:
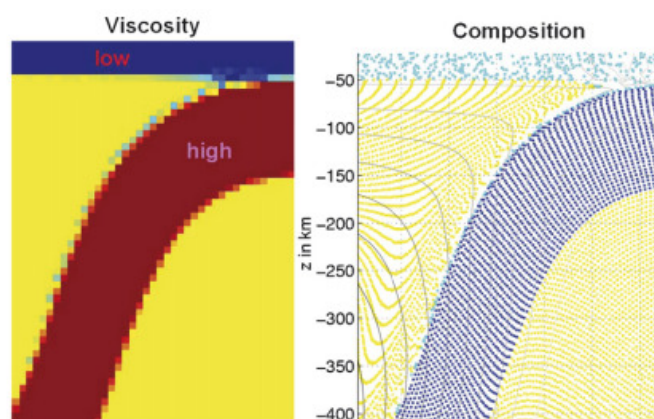
Taken from Schmeling *et al.* (2008) [1200]. Details of the entrainment and lubrication of the soft surface layer. Light blue particles are sticky air particle and are found to greatly alter the viscosity of the subduction channel.

4. it is not clear how thick the air layer must be

5. it often requires to ascribe thermal parameters to the air;

---

[1]Sometimes a zero value is problematic and one then resorts to a very low value instead.
[2]https://en.wikipedia.org/wiki/Viscosity

6. it makes the implementation of Dirichlet or Neuman boundary conditions for temperature at the surface less obvious.

7. it makes the coupling with surface processes codes less straightforward.

8. its accuracy depends on the method used to track materials in the rest of the code (markers, level sets, ...). If markers are used, the free surface position is then known up to the average distance between markers.

9. it negatively impacts the condition number of the matrix.

The term 'sticky water' is sometimes employed too. The dynamic viscosity of water is about 10^{-3} Pa s so that it is also negligible compared to the viscosity of Earth materials and the same reasonng as air applies. However, in such a case a density of about 1000 kg m^{-3} is then assigned to the layer (e.g. Gerya & Burg (2007) [479]).

In conclusion, as stated in Crameri *et al.* (2012) [304]: "the sticky air method is a good way to simulate a free surface for Eulerian approaches, provided that its parameters are chosen carefully."

The Arbitrary Lagrangian Eulerian (ALE) approach

It is a very widely used approach in FEM-based geodynamics codes but originates in the field of CFD (Hirt *et al.* (1974)[608], Hughes *et al.* (1981) [642]) and is described at length in Souli & Zolesio (2001) [1262], Donea *et al.* (2004) [361], Donea & Huerta [363]. To put it very simply, the key idea in the ALE formulation is the introduction of a computational mesh which can move and deform with a velocity independent of the velocity carried by the material particles.

The simple approach in Thieulot (2011) [1343]. What follows is written with a 2D Cartesian model in mind ($Q_1 \times P_0$ elements are used). The computational domain is a rectangle of size $L_x \times L_y$ and a nnx $\times$ nny rectangular grid spanning the simulation domain is generated. The grid points constituting the top row of the grid define the discrete free surface of the domain. Once the Eulerian velocity field has been computed on these, their position is first updated using a simple Eulerian advection step (see a,b on figure hereunder):

$$\vec{r}'_i(t + \delta t) = \vec{r}_i(t) + \vec{v}_i \cdot \delta t \qquad i = 1, \dots nnx$$

The other boundaries of the system remain fixed at locations $x = 0$, $x = L_x$ and $y = 0$. Even though the Eulerian grid must conform to the current domain shape, only vertical motion of grid nodes is allowed. It is therefore necessary to resample the predicted free surface given by $\vec{r}'_i$ at equidistant positions between $x = 0$ and $x = L_x$. The resampling is carried out either with Spline functions or a moving least square algorithm. Finally, the vertical position of all the nodes corresponding to column $i \in [1, nnx]$ is recalculated so that they are equidistant, as sketched in Figure d. This has the advantage of keeping the mesh distortion to a minimum in the case of large deformation.
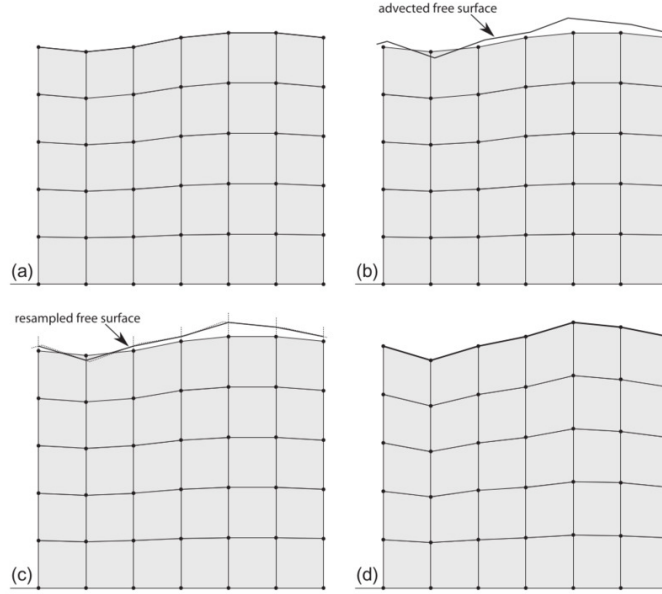
The ALE algorithm of [1343] in 2D. (a) Grid and free surface at a given time t; (b) advection of the free surface; (c) resampling of the free surface at equidistant abscissae; (d) vertical adjustment of grid nodes in each column at equidistant ordinates.

The ALE method is used in the SOPALE , SULEC , FANTOM , ELEFANT , and ASPECT codes to name a few (see Appendix **??**) although each code has its own specific implementation, as detailed in what follows for ASPECT .

The not-so-simple but rather elegant approach of Aspect   What follows is mostly borrowed from Rose *et al.* [1157]. Their approach has the advantage that it does not presuppose a geometry (Cartesian, Spherical, ...) nor a number of dimensions. It is also designed to work in parallel and on octree-based meshes, and with various combinations of boundary conditions. Note that the authors specify that "for moderate mesh deformation, the mesh stays smooth and well conditioned, though it breaks down for large deformations".

This approach is obtained by simply imposing the obvious condition that no particle (fluid parcel) can cross the free surface (because it is a material surface). This can be imposed in a straightforward manner by using a Lagrangian description along this surface. However, this condition may be relaxed by imposing only the necessary condition: $\vec{v}$ equal to zero along the normal to the boundary (ie. $\vec{n} \cdot \vec{v} = 0$, where $\vec{n}$ is the outward unit normal to the fluid domain). The mesh position, normal to the free surface, is determined from the normal component of the particle velocity and remeshing can be performed along the tangent; see, for instance Huerta and Liu (1988) [636] or Braess and Wriggers (2000) [138].

As mentioned above, the mesh velocity in normal direction at the free surface (with unit normal $\vec{n}$) has to be consistent with the velocity of the Stokes velocity solution $\vec{v}(t)$:

$$\vec{v}_{\text{mesh}}(t) \cdot \vec{n} = \vec{v}(t) \cdot \vec{n} \qquad \text{on} \quad \Gamma_F \tag{16.1}$$

In ALE calculations the internal mesh velocity is usually undetermined, but one wants to smoothly deform the mesh so as to preserve its regularity, avoiding inverted or otherwise poorly conditioned cells. The mesh deformation can be calculated in many different ways, icluding algebraic (as mentioned in the previous paragraph) and PDE based approaches. The latter is chosen here. The Laplace equation is solved where the unknown is the mesh velocity, i.e. one must solve:

$$\Delta \vec{v}_{\text{mesh}} = 0 \tag{16.2}$$

subjected to the following boundary conditions:

$$\begin{aligned}
\vec{v}_{\text{mesh}} &= \vec{0} & \text{on } \Gamma_0 \\
\vec{v}_{\text{mesh}} &= (\vec{v} \cdot \vec{n})\vec{n} & \text{on } \Gamma_F \\
\vec{v}_{\text{mesh}} \cdot \vec{n} &= 0 & \text{on } \Gamma_{FS}
\end{aligned}$$
(16.3)

where Γ_{FS} is the part of the boundary with free slip boundary conditions, Γ_0 is the no-slip part and Γ_{FS} is the free slip part.

Once the mesh velocity has been obtained for all mesh points, these can be moved with said velocity. However, it must be noted that the multiple occurences of the normal vector in the above equations is not without problem as the normal vectors are not well defined on the mesh vertices, which is where the mesh velocity is defined.

`INSERT FIGURE`

This yields what the authors coin the 'quasi-implicit' scheme (we have so far neglected any kind of stabilisation):

1. Solve the Stokes system;

2. Solve for the surface mesh velocity using Equation 16.4;

3. Solve for the internal mesh velocity using Equations 16.2, 16.3;

4. Advect the mesh forward in time using displacements determined by the forward Euler scheme: $\vec{x}(t^{n+1}) = \vec{x}(t^n) + \vec{v}_{\text{mesh}}\delta t$.

The authors list two simple methods of computing the normals:

- one can take $\vec{n}$ as the direction of the local vertical,

- one could compute $\vec{n}$ as some weighted average of the cell normals adjacent to a given vertex

but conclude that they have found that these schemes do not necessarily have good mass conservation properties.

A better approach is proposed in the form of an L_2 projection of the normal velocity $\vec{v} \cdot \vec{n}$ onto the free surface Γ_F. Multiplying the boundary conditions $\vec{v}_{\text{mesh}} = (\vec{v} \cdot \vec{n})\vec{n}$ by a test function $\vec{w}$ and integrating over the free surface part of the boundary, we find:

$$\int_{\Gamma_F} \vec{w} \cdot \vec{v}_{\text{mesh}} d\Gamma = \int_{\Gamma_F} \vec{w} \cdot (\vec{v} \cdot \vec{n})\vec{n} d\Gamma = \int_{\Gamma_F} (\vec{w} \cdot \vec{n})(\vec{v} \cdot \vec{n}) d\Gamma$$
(16.4)

When discretized, this forms a linear system which can be solved for the mesh velocity $\vec{v}_{\text{mesh}}$ at the free surface. This system, being nonzero over only the free surface, is relatively computationally inexpensive to solve. The authors unfortunately fail to mention that this approach is particularly interesting since the numerical quadrature used to compute the above integrals require the normal $\vec{n}$ between the nodes and these normals are well defined over each segment joining two nodes![3]

In what follows I present in some detail how to carry out the L_2 projection to arrive at the surface velocity for both Q_1 and Q_2 elements.

I start from the following integral over a Q_1 element:

$$\int_{\Gamma_e} N_i \vec{v}_{mesh} d\Gamma = \int N_i \begin{pmatrix} u_{mesh} \\ v_{mesh} \end{pmatrix} d\Gamma$$
(16.5)

$$= \int N_i \begin{pmatrix} N_1 & 0 & N_2 & 0 \\ 0 & N_1 & 0 & N_2 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{pmatrix} d\Gamma$$
(16.6)

(16.7)

---

[3]what if Q_k with $k > 1$ elements are used and nodes on the surface no more form a line?

Writing this equation alternatively for $N_i = N_1, N_2$ yields:

$$\int_{\Gamma_e} \begin{pmatrix} N_1 N_1 & 0 & N_1 N_2 & 0 \\ 0 & N_1 N_1 & 0 & N_1 N_2 \\ N_2 N_1 & 0 & N_2 N_2 & 0 \\ 0 & N_2 N_1 & 0 & N_2 N_2 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{pmatrix} d\Gamma = \int_{\Gamma_e} \begin{pmatrix} N_1 N_1 & 0 & N_1 N_2 & 0 \\ 0 & N_1 N_1 & 0 & N_1 N_2 \\ N_2 N_1 & 0 & N_2 N_2 & 0 \\ 0 & N_2 N_1 & 0 & N_2 N_2 \end{pmatrix} d\Gamma \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{pmatrix}$$

Turning now to the right hand side $\int_{\Gamma_e} N_i (\vec{v} \cdot \vec{n}_e) \vec{n}_e d\Gamma$, it yields the following rhs:

$$\int_{\Gamma_e} (\vec{v} \cdot \vec{n}_e) \begin{pmatrix} N_1 n_x \\ N_1 n_y \\ N_2 n_x \\ N_2 n_y \end{pmatrix} d\Gamma$$

The elemental matrix and rhs must be built for each element and assembled in a global matrix and rhs. The solution is the mesh velocity vector at all surface nodes. the same approach can be taken for Q_2 elements:

$$\int_{\Gamma_e} N_i \vec{v}_{mesh} d\Gamma \;=\; \int N_i \begin{pmatrix} u_{mesh} \\ v_{mesh} \end{pmatrix} d\Gamma \tag{16.8}$$

$$= \int N_i \begin{pmatrix} N_1 & 0 & N_2 & 0 & N_3 & 0 \\ 0 & N_1 & 0 & N_2 & 0 & N_3 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \end{pmatrix} d\Gamma \tag{16.9}$$

Writing this equation alternatively for $N_i = N_1, N_2, N_3$ yields:

$$\int_{\Gamma_e} \begin{pmatrix} N_1 N_1 & 0 & N_1 N_2 & 0 & N_1 N_3 & 0 \\ 0 & N_1 N_1 & 0 & N_1 N_2 & 0 & N_1 N_3 \\ N_2 N_1 & 0 & N_2 N_2 & 0 & N_2 N_3 & 0 \\ 0 & N_2 N_1 & 0 & N_2 N_2 & 0 & N_2 N_3 \\ N_3 N_1 & 0 & N_3 N_2 & 0 & N_3 N_3 & 0 \\ 0 & N_3 N_1 & 0 & N_3 N_2 & 0 & N_3 N_3 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \end{pmatrix} d\Gamma$$

$$= \int_{\Gamma_e} \begin{pmatrix} N_1 N_1 & 0 & N_1 N_2 & 0 & N_1 N_3 & 0 \\ 0 & N_1 N_1 & 0 & N_1 N_2 & 0 & N_1 N_3 \\ N_2 N_1 & 0 & N_2 N_2 & 0 & N_2 N_3 & 0 \\ 0 & N_2 N_1 & 0 & N_2 N_2 & 0 & N_2 N_3 \\ N_3 N_1 & 0 & N_3 N_2 & 0 & N_3 N_3 & 0 \\ 0 & N_3 N_1 & 0 & N_3 N_2 & 0 & N_3 N_3 \end{pmatrix} d\Gamma \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \end{pmatrix} \tag{16.10}$$

The right hand side is then

$$\int_{\Gamma_e} (\vec{v} \cdot \vec{n}_e) \begin{pmatrix} N_1 n_x \\ N_1 n_y \\ N_2 n_x \\ N_2 n_y \\ N_3 n_x \\ N_3 n_y \end{pmatrix} d\Gamma$$

Having obtained the boundary condition velocity for the Laplace equation, we can now turn our attention to solving this ODE.

In what follows I omit the subscript 'mesh' and focus on the 2D case. The components of the (mesh) velocity are given by

$$u^h = \sum_{i=1}^{m_v} N_i^v u_i \qquad\qquad v^h = \sum_{i=1}^{m_v} N_i^v v_i \qquad\qquad \vec{v}^h = \begin{pmatrix} u^h \\ v^h \end{pmatrix}$$

We start from the ODE to solve in its strong form:

$$\Delta \vec{v}^h = \vec{0}$$

We multiply it by a velocity test function N_i^v and integrate over an element:

$$
\begin{aligned}
& \vec{0} \\
=\ & \int_{\Omega_e} N_i^v \Delta \vec{v}^h \\
=\ & \int_{\Omega_e} N_i^v \Delta \vec{v}^h \, dV \\
=\ & \int_{\Omega_e} \begin{pmatrix} N_i^v \Delta u^h \\ N_i^v \Delta v^h \end{pmatrix} dV \\
=\ & \int_{\Omega_e} \begin{pmatrix} N_i^v \vec{\nabla} \cdot \vec{\nabla} u^h \\ N_i^v \vec{\nabla} \cdot \vec{\nabla} v^h \end{pmatrix} dV \\
=\ & \int_{\Omega_e} \begin{pmatrix} \vec{\nabla} N_i^v \cdot \vec{\nabla} u^h \\ \vec{\nabla} N_i^v \cdot \vec{\nabla} v^h \end{pmatrix} dV \\
=\ & \int_{\Omega_e} \begin{pmatrix} \partial_x N_i^v \partial_x u^h + \partial_y N_i^v \partial_y u^h \\ \partial_x N_i^v \partial_x v^h + \partial_y N_i^v \partial_y v^h \end{pmatrix} dV \\
=\ & \int_{\Omega_e} \begin{pmatrix} \partial_x N_i^v & \partial_y N_i^v & 0 & 0 \\ 0 & 0 & \partial_x N_i^v & \partial_y N_i^v \end{pmatrix} \cdot \begin{pmatrix} \partial_x u^h \\ \partial_y u^h \\ \partial_x v^h \\ \partial_y v^h \end{pmatrix} dV
\end{aligned}
$$

$$
=\ \int_{\Omega_e} \begin{pmatrix} \frac{\partial N_i^v}{\partial x} & \frac{\partial N_i^v}{\partial y} & 0 & 0 \\ 0 & 0 & \frac{\partial N_i^v}{\partial x} & \frac{\partial N_i^v}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial N_1^v}{\partial x} & 0 & \frac{\partial N_2^v}{\partial x} & 0 & \cdots & \frac{\partial N_{m_v}^v}{\partial x} & 0 \\ \frac{\partial N_1^v}{\partial y} & 0 & \frac{\partial N_2^v}{\partial y} & 0 & \cdots & \frac{\partial N_{m_v}^v}{\partial y} & 0 \\ 0 & \frac{\partial N_1^v}{\partial x} & 0 & \frac{\partial N_2^v}{\partial x} & \cdots & 0 & \frac{\partial N_{m_v}^v}{\partial x} \\ 0 & \frac{\partial N_1^v}{\partial y} & 0 & \frac{\partial N_2^v}{\partial y} & \cdots & 0 & \frac{\partial N_{m_v}^v}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \cdots \\ u_{m_v} \\ v_{m_v} \end{pmatrix} dV
$$

Writing this equation for $i = 1, \dots m_v$, we obtain:

$$
\int \begin{pmatrix} \frac{\partial N_1^v}{\partial x} & \frac{\partial N_1^v}{\partial y} & 0 & 0 \\ 0 & 0 & \frac{\partial N_1^v}{\partial x} & \frac{\partial N_1^v}{\partial y} \\ \frac{\partial N_2^v}{\partial x} & \frac{\partial N_2^v}{\partial y} & 0 & 0 \\ 0 & 0 & \frac{\partial N_2^v}{\partial x} & \frac{\partial N_2^v}{\partial y} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial N_{m_v}^v}{\partial x} & \frac{\partial N_{m_v}^v}{\partial y} & 0 & 0 \\ 0 & 0 & \frac{\partial N_{m_v}^v}{\partial x} & \frac{\partial N_{m_v}^v}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial N_1^v}{\partial x} & 0 & \frac{\partial N_2^v}{\partial x} & 0 & \cdots & \frac{\partial N_{m_v}^v}{\partial x} & 0 \\ \frac{\partial N_1^v}{\partial y} & 0 & \frac{\partial N_2^v}{\partial y} & 0 & \cdots & \frac{\partial N_{m_v}^v}{\partial y} & 0 \\ 0 & \frac{\partial N_1^v}{\partial x} & 0 & \frac{\partial N_2^v}{\partial x} & \cdots & 0 & \frac{\partial N_{m_v}^v}{\partial x} \\ 0 & \frac{\partial N_1^v}{\partial y} & 0 & \frac{\partial N_2^v}{\partial y} & \cdots & 0 & \frac{\partial N_{m_v}^v}{\partial y} \end{pmatrix} \cdot \underbrace{\begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \cdots \\ u_{m_v} \\ v_{m_v} \end{pmatrix}}_{\vec{V}} dV = \vec{0}
$$

or,

$$\left(\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{B} \; dV \right) \cdot \vec{V} = \vec{0}$$

where $\boldsymbol{B}$ is a $(ndim * ndim) \times (m_v * ndofV)$ matrix. This is implemented in Stone 54 **??**.

Remark. *The integration by parts should have a minus appear but since the left hand side is 0, it is not taken into account.*

Note that Rose *et al.* (2017) [1157] go further than this and propose a 'nonstandard finite difference scheme' and make a link with the stabilisation presented in Kaus *et al.* (2010) [722].

surface terms arising from the integration by parts are neglected. EXPLAIN WHY!

Yet another approach [361] The unknown position of free surfaces can be computed using the following approach: for the simple case of a single-valued function $h = h(x, y, t)$, a hyperbolic equation must be solved,

$$\frac{\partial h}{\partial t} + (\vec{v} \cdot \vec{\nabla})h = 0 \tag{16.11}$$

This is the kinematic equation of the surface and has been used, for instance, by Ramaswamy and Kawahara (1987), Huerta and Liu, 1988b, 1990; Souli and Zolesio (2001).

Eq. (16.11) is a simple advection equation. One could also add a diffusion operator with a diffusion coefficient D. Low values of D could be used to stabilise the surface while higher values (possibly nonlinear ones) could be used to account for simple surface processes.

$$\frac{\partial h}{\partial t} + (\vec{v} \cdot \vec{\nabla})h = D\Delta h \tag{16.12}$$

Also, Hansen and Nielsen [566, 565] write: "During the entire model evolution surface processes act to re-distribute sediments. These processes are modelled by a diffusion equation with a source term enabling the transport of sediments to and from the model profile. The transport equation is written

$$\dot{h} = \nabla \cdot (\kappa \nabla h) + \dot{s}(w)$$

where $\kappa = 200 km^2/Ma$ is the diffusivity of topography and $\dot{s}(w)$ is a linear function of water depth."

The following pictures are taken from Naliboff *et al.* (2017) [982] on the topic of how complex fault interaction controls continental rifting. It is a beautiful example (among many) of the importance of free surface geodynamical expression and large deformation:
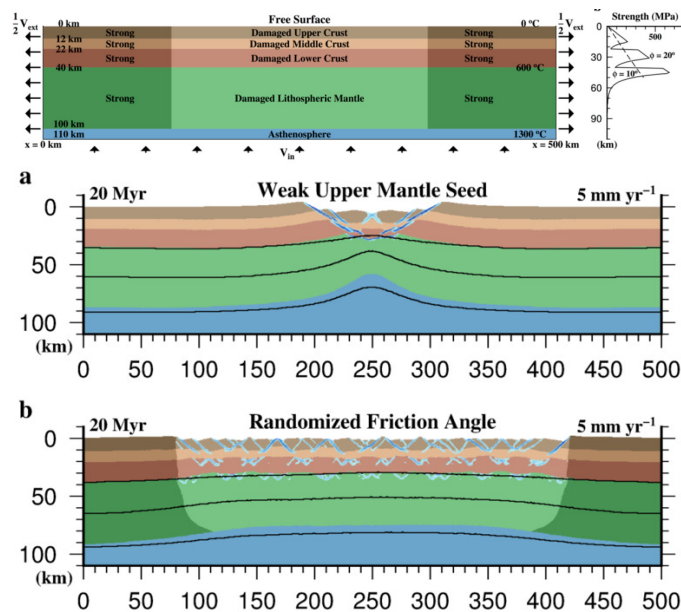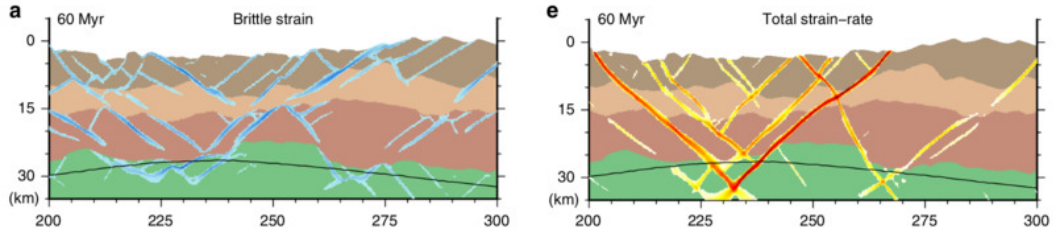
On the topic of moving internal nodes

Braess & Wriggers [138] propose the following interesting algorithm: "A measure of the quality of a triangular mesh is the quotient of the outer radius r_{out} and the inner radius r_{in} of each element. This quotient is important because it plays a certain role in a priori error estimates. If an element degenerates this quotient will approach infinity. Another important feature of good mesh is that no element becomes very large. With these considerations in mind the penalty function W is defined:

$$W = \sum_{elts} \left(\frac{r_{out}}{r_{in}} \right)^m \left(\frac{r_{out}}{r_0} \right)^n \tag{16.13}$$

where m, n and r_0 are positive constants. For our calculations we chose $m = 3$, $n = 1$ and $r_0 = 1$, but the results seem to depend only slightly on this choice. Whenever a triangle is distorted or very large, this function becomes very large. A similar penalty function was presented in [695] for four-node elements. In that case the angles of the elements are used to construct the penalty function. In order to regularize a distorted mesh the coordinates of the internal nodes will be chosen such that W is minimized. It is not necessary to reach the global minimum, a rough approximation is sufficient. Therefore the minimization of the potential can be done efficiently with standard procedures and will not be discussed in any detail. This algorithm can also be applied to h-adaptive mesh-generation by choosing appropriate constants r_0 for each triangle."[4]

This is still WORK IN PROGRESS. I Need to look at those papers:

ALE: Ramaswamy & Kawahara (1987) [1101], ALE: Huerta & Liu (1988) [636], ALE: Ponthot & Belytschko (1998)[1075], ALE: Benson (1989) [76], ALE: Sung et al. (2000) [1303], ALE: Ramaswamy (1990) [1100], Tezduyar et al. (1992) [1334](moving pulse) Andres-Martinez et al. (2015) [24] (read !! extract benchmarks ?) Kramer et al. (2012) [775] (read !! extract benchmarks ?) Steer et al. (2011) [1276] Maierova [871] Zhong et al. [1519], Gurnis et al. [546] Ellis et al. (2004) [392] Braess & Wriggers (2000) [138] Parsons & Daly (1983) [1040]

Free surface treatment in FDM: [376].

Free surface stabilization algorithm (FSSA)

To start with, Duretz et al. (2011) [375] is concerned with the Finite Difference method so I will not expore this one further (see Gerya's book [484] for all things FDM). Note that their approach is similar to what follows.

---

[4]Indeed, if r_0 is the same for all elements this parameter will not play any role at all in the minimisation process.

There are then four articles left: Kaus *et al.* (2010) [722], Quinquis & Buiter (2011) [1094], Rose *et al.* (2017) [1157], and Schuh-Senlis *et al.* (2020) [1221].

The first one is rather technical while the second is more to the point with some simple reasoning. Apparently both author groups arrived at about the same formulation/algorithm at about the same time. The third and fourth paper implement the algorithm in ASPECT and the FAIStokes code respectively.

The basic idea is rather simple. As stated in Quinquis *et al.* (2011) [1094]: "In numerical subduction models which include a free surface or other interfaces with large density contrasts, an instability can occur as a result of numerical overshoot when computing restoring forces".

In essence when the timestep δt exceeds (or is comparable) to a characteristic relaxation time t_s then body and surface forces will be out of balance with the updated free surface. As a consequence oscillations will occur and get amplified with sometimes a sloshing effect, also called 'drunken sailor'. . The solution is to add a term in the FE matrix which takes into account the incremental change in normal forces across density interfaces during each timestep:

$$\Delta F_y = \Delta \rho \ g_y \delta t \ v_y$$

where ΔF_y is the extra vertical force term across the interface, $\Delta \rho$ is the change in density across the interface, and v_y is the vertical velocity on the interface. This extra term is applied to the free surface, and at any other interface with a density contrast. Unfortunately, although this is easy to understand, it is not very clear how to implement it, which is when Kaus *et al.* (2010) becomes useful.

I first present results from my doomed 2014 ELEFANT paper [5] [1342] in which I carried out the Rayleigh-Taylor instability experiment described in [722] where a dense fluid overlays another fluid in a square domain with their initial interface being of sinusoidal shape.
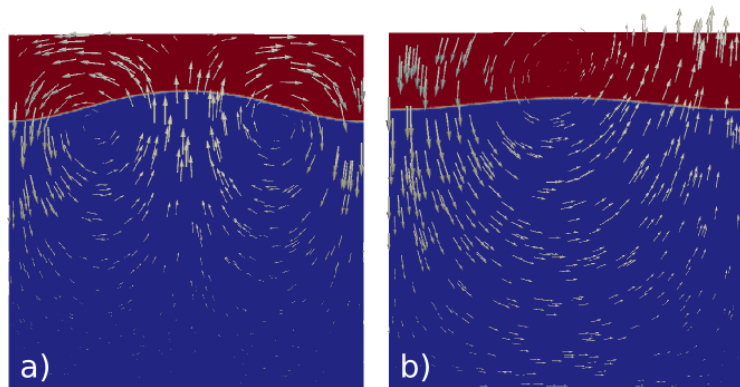
Taken from Thieulot (2014) [1342] When the time step δt is small, the simulation evolves smoothly, but too large a time step leads to a sloshing instability,

also called 'drunken sailor effect' in the community.

The position $y(t)$ of the free surface point situated at $x = L_x$ is monitored over time:

---

[5]This paper is about my ELEFANT code which relied at the time solely on $Q_1 \times P_0$ elements with the penalty formulation.
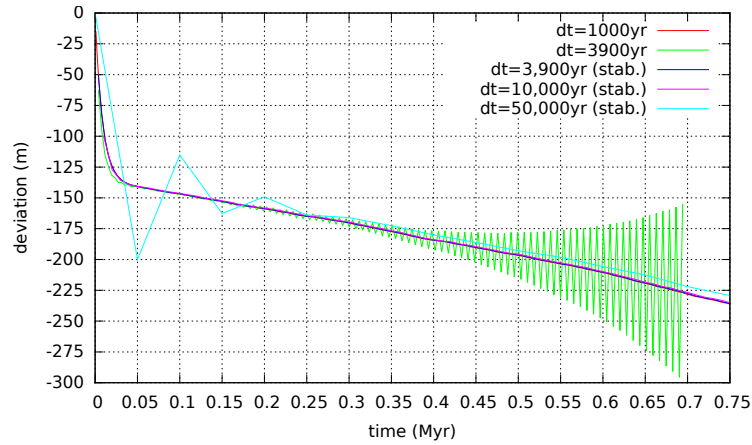
The stabilisation algorithm is first switched off and up to a time step of $\delta t \sim 3800$yr, the free surface does not develop any instability (all curves are superimposed). For a time step of 3900yr and above, a clear oscillation develops (green seesaw curve on the figure). Turning the algorithm on, time steps larger than $dt \sim 10,000$yr can be used and no oscillation is observed.

I found that: a) the presence of the stabilisation algorithm with small time steps does not introduce a significant difference in the outcome (less than a meter of deviation after 1Myr); b) the time step was increased up to $\delta t = 20,000$yr and the simulation remained stable (the vertical deviation differed by approximately 4m from the one obtained with a very small time step, which remains a remarkable result since it only represents $\sim 0.2\%$ of the element size); c) time steps up to $\delta t = 50,000$yr remain stable but lead to oscillations at the beginning and deviate in the end by about 5 meters.

Finally, let us look at the mathematical details behind this stabilisation, as presented in Kaus *et al.* (2010) [722].

FINISH!!

📦 Relevant Literature: Furuchi (2011) [454]

16.2 Picard and Newton

16.2.1 Picard iterations

Let us consider the following system of nonlinear algebraic equations:

$$\mathbb{A}(\vec{X}) \cdot \vec{X} = \vec{b}(\vec{X})$$

Both matrix and right hand side depend on the solution vector $\vec{X}$.

For many mildly nonlinear problems, a simple successive substitution iteration scheme (also called Picard method) will converge to the solution and it is given by the simple relationship:

$$\mathbb{A}(\vec{X}^n) \cdot \vec{X}^{n+1} = \vec{b}(\vec{X}^n)$$

where n is the iteration number. It is easy to implement:

1. guess $\vec{X}^0$ or use the solution from previous time step

2. compute $\mathbb{A}$ and $\vec{b}$ with current solution vector $\vec{X}^{old}$

3. solve system, obtain T^{new}

4. check for convergence (are $\vec{X}^{old}$ and $\vec{X}^{new}$ close enough?)

5. $\vec{X}^{old} \leftarrow \vec{X}^{new}$

6. go back to 2.

There are various ways to test whether iterations have converged. The simplest one is to look at $\left\| \vec{X}^{old} - \vec{X}^{new} \right\|$ (in the L_1, L_2 or maximum norm) and assess whether this term is smaller than a given tolerance ϵ. However this approach poses a problem: in geodynamics, if two consecutively obtained temperatures do not change by more than a thousandth of a Kelvin (say $\epsilon = 10^{-3}$K) we could consider that iterations have converged but looking now at velocities which are of the order of a cm/year (i.e. $\sim 3 \cdot 10^{-11}$m/s) we would need a tolerance probably less than 10^{-13}m/s. We see that using absolute values for a convergence criterion is a potentially dangerous affair, which is why one uses a relative formulation (thereby making ϵ a dimensionless parameter):

$$\frac{\left\| \vec{X}^{old} - \vec{X}^{new} \right\|}{\left\| \vec{X}^{new} \right\|} < \epsilon$$

Another convergence criterion is proposed by Reddy (section 3.7.2) [1118]:

$$\left(\frac{(\vec{X}^{old} - \vec{X}^{new}) \cdot (\vec{X}^{old} - \vec{X}^{new})}{X^{new} \cdot X^{new}} \right)^{1/2} < \epsilon$$

Yet another convergence criterion is used in [1343]: the means $< \vec{X}^{old} >$, $< \vec{X}^{new} >$ as well as the variances σ^{old} and σ^{new} are computed, followed by the correlation factor R:

$$R = \frac{< (\vec{X}^{old} - < \vec{X}^{old} >) \cdot (\vec{X}^{new} - < \vec{X}^{new} >) >}{\sqrt{\sigma^{old} \sigma^{new}}}$$

Since the correlation is normalised, it takes values between 0 (very dissimilar velocity fields) and 1 (very similar fields). The following convergence criterion is then used: $1 - R < \epsilon$.

Note that in some instances and improvement in convergence rate can be obtained by use of a relaxation formula where one first solves

$$\mathbb{A}(\vec{X}^n) \cdot \vec{X}^\star = \vec{b}(\vec{X}^n)$$

and then updates $\vec{X}^n$ as follows:

$$\vec{X}^n = \gamma \vec{X}^n + (1 - \gamma)\vec{X}^\star \qquad 0 < \gamma \le 1$$

When $\gamma = 1$ we recover the standard Picard iterations formula above.

16.2.2   Defect correction formulation

Work in progress.

We start from the system to solve:

$$\boldsymbol{A}(\vec{X}) \cdot \vec{X} = \vec{b}(\vec{X})$$

with the associated residual vector $\vec{F}$

$$\vec{F}(\vec{X}) = \boldsymbol{A}(\vec{X}) \cdot \vec{X} - \vec{b}(\vec{X})$$

The Newton-Raphson algorithm consists of two steps:

1. solve $\boldsymbol{J}_k \cdot \delta \vec{X}_k = -\vec{F}(\vec{X}_k)$, or in the case of the incompressible Stokes equation FEM system:

$$\begin{pmatrix} \boldsymbol{J}_k^{\mathcal{V}\mathcal{V}} & \boldsymbol{J}_k^{\mathcal{V}\mathcal{P}} \\ \boldsymbol{J}_k^{\mathcal{P}\mathcal{V}} & 0 \end{pmatrix} \cdot \begin{pmatrix} \delta\vec{\mathcal{V}}_k \\ \delta\vec{\mathcal{P}}_k \end{pmatrix} = \begin{pmatrix} -\vec{F}_k^{\mathcal{V}} \\ -\vec{F}_k^{\mathcal{P}} \end{pmatrix}$$

2. update $\vec{X}_{k+1} = \vec{X}_k + \alpha_k \delta\vec{X}_k$

The defect correction Picard approach consists of neglecting the derivative terms present in the J terms (Eqs. 16,17,18 of [441]) so that

$$\boldsymbol{J}_k^{\mathcal{V}\mathcal{V}} \simeq \mathbb{K}_k \qquad \boldsymbol{J}_k^{\mathcal{V}\mathcal{P}} \simeq \mathbb{G} \qquad \boldsymbol{J}_k^{\mathcal{P}\mathcal{V}} \simeq \mathbb{G}^T$$

and step 1 of the above iterations become:

$$\begin{pmatrix} \mathbb{K}_k & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \delta\vec{\mathcal{V}}_k \\ \delta\vec{\mathcal{P}}_k \end{pmatrix} = \begin{pmatrix} -\vec{F}_k^{\mathcal{V}} \\ -\vec{F}_k^{\mathcal{P}} \end{pmatrix}$$

🪨 Relevant Literature: Quasi Newton methods [396]. Also check Christensen and Yuen [264] for a succinct N-R explanation in the context of stream fct formulation.

- VV Ermakov and Nikolai Nikolaevich Kalitkin. "The optimal step and regularization for Newton's method". In: *USSR Computational Mathematics and Mathematical Physics* 21.2 (1981), pp. 235–242. DOI: 10.1016/0041-5553(81)90022-7

- MS Engelman, Gilbert Strang, and K-J Bathe. "The application of quasi-Newton methods in fluid mechanics". In: *Int. J. Num. Meth. Eng.* 17.5 (1981), pp. 707–718

- Dana A Knoll and David E Keyes. "Jacobian-free Newton–Krylov methods: a survey of approaches and applications". In: *J. Comp. Phys.* 193.2 (2004), pp. 357–397. DOI: `10.1016/j.jcp.2003.08.010`

- He Yiqian and Yang Haitian. "Solving inverse couple-stress problems via an element-free Galerkin (EFG) method and Gauss–Newton algorithm". In: *Finite Elements in Analysis and Design* 46.3 (2010), pp. 257–264. DOI: `10.1016/j.finel.2009.09.009`

- Pierre Saramito. "A damped Newton algorithm for computing viscoplastic fluid flows". In: *Journal of Non-Newtonian fluid mechanics* 238 (2016), pp. 6–15. DOI: `10.1016/j.jnnfm.2016.05.007`

- M.R.T. Fraters, W. Bangerth, C. Thieulot, A.C. Glerum, and W. Spakman. "Efficient and Practical Newton Solvers for Nonlinear Stokes Systems in Geodynamic Problems". In: *Geophy. J. Int.* 218 (2019), pp. 873–894. DOI: `10.1093/gji/ggz183`

- Johann Rudi, Yu-hsuan Shih, and Georg Stadler. "Advanced Newton methods for geodynamical models of Stokes flow with viscoplastic rheologies". In: *Geochem. Geophys. Geosyst.* 21.9 (2020), e2020GC009059. DOI: `10.1029/2020GC009059`

16.3 Convergence criterion for nonlinear iterations

Disclaimer: the topic of nonlinear PDEs solving is vast and has received much attention from the mathematical community. In what follows I present a few key ideas which are at the core of many codes and publications in computational geodynamics.

Looking at the conservation equations that we must solve, i.e. conservation of mass, momentum and energy, we find that more often than not the coefficients of these PDEs depend on the strain rate, temperature, pressure, etc ... This makes solving the PDEs even harder. Also the advection term $\vec{v} \cdot \vec{\nabla}$ couples the two primary variables velocity and temperature.

One simple approach consists first in 'separating' the mass+momentum equations from the energy equation: one solves the first two equations assuming temperature known while the energy equation is solved assuming velocity and pressure known. Better schemes obviously exist and iterate on these equations until convergence for velocity, pressure and temperature is reached (see for instance the ASPECT manual). In what follows I focus on the mass and momentum equations assuming temperature known.

The main source of nonlinearity lies in the (effective) viscosity which often depends on strain rate and pressure (note that density can also depend on pressure in compressible cases):

$$\vec{\nabla} \cdot (2\eta_{\text{eff}}(\dot{\boldsymbol{\varepsilon}}, p) \, \dot{\boldsymbol{\varepsilon}}) - \vec{\nabla}p + \rho\vec{g} = \vec{0}$$
$$\vec{\nabla} \cdot \vec{v} = 0$$

Simply put, in order to solve these equations and obtain the velocity and pressure fields I need to specify the density and viscosity (and of course appropriate boundary conditions!), but in order to compute the viscosity I need the strain rate and pressure fields.

The simplest approach here consists in so-called Picard iterations as explained in Section 16.2.1.

Let us start with the penalty-based FEM codes. In this case the mass and momentum equations are 'merged' into a single PDE where pressure has been eliminated:

$$\vec{\nabla} \cdot (2\eta_{\text{eff}}(\dot{\boldsymbol{\varepsilon}}, p) \, \dot{\boldsymbol{\varepsilon}}) + \lambda\vec{\nabla}(\vec{\nabla} \cdot \vec{v}) + \rho\vec{g} = \vec{0}$$

In this case the algorithm is simple:

1. start with a guess for the velocity and pressure fields, i.e. $\vec{\mathcal{V}}^{old}$ and $\vec{\mathcal{P}}^{old}$

2. compute the effective viscosity field with $\vec{\mathcal{V}}^{old}$ and $\vec{\mathcal{P}}^{old}$

3. solve PDE, obtain new solution $\vec{\mathcal{V}}^{new}$

4. compute $\vec{\mathcal{P}}^{new}$ from $\vec{\mathcal{V}}^{new}$

5. assess convergence, i.e. answer 'how close are the newly obtained fields from the old ones?'

6. $\vec{\mathcal{V}}^{old} \leftarrow \vec{\mathcal{V}}^{new}$, and $\vec{\mathcal{P}}^{old} \leftarrow \vec{\mathcal{P}}^{new}$

7. if not converged go back to 2, else exit

Thieulot (2011) [1343] computes the means $\langle \vec{\mathcal{V}}^i \rangle$, $\langle \vec{\mathcal{V}}^{i+1} \rangle$, and the variances $\sigma_{\mathcal{V}}^i$ and $\sigma_{\mathcal{V}}^{i+1}$ followed by the correlation

$$R^{i,i+1} = \frac{\langle (\vec{\mathcal{V}}^i - \langle \vec{\mathcal{V}}^i \rangle) \cdot (\vec{\mathcal{V}}^{i+1} - \langle \vec{\mathcal{V}}^{i+1} \rangle) \rangle}{\sqrt{\sigma_{\mathcal{V}}^i \sigma_{\mathcal{V}}^{i+1}}}$$

Since the correlation is normalised, it takes values between 0 (very dissimilar velocity fields) and 1 (very similar fields). The following convergence criterion, formulated in terms of the variable $\chi = 1 - R^{i,i+1}$ has been implemented: convergence is reached when $\chi < tol$. Since pressure is a derived quantity from velocity, if velocity is converged so is pressure[6].

When the algorithm above is close to convergence then $\vec{\mathcal{V}}^i$ and $\vec{\mathcal{V}}^{i-1}$ are close. If these were scalar quantities we could subtract them and look at the (absolute) difference: if it is 'small enough' then the algorithm has converged. However there are two problems with this:

1. $\vec{\mathcal{V}}^i$ and $\vec{\mathcal{V}}^{i-1}$ are vector quantities (with potentially millions of values) so in order to measure a scalar difference between these we must take the norm of the difference, or $||\vec{\mathcal{V}}^i - \vec{\mathcal{V}}^{i-1}||$ and it is common to take the L^2-norm.

2. we do not know a priori the (magnitude of the) solution so that 'small enough' is a dangerous statement. We could check for $||\vec{\mathcal{V}}^i - \vec{\mathcal{V}}^{i-1}|| < tol$ and set $tol = 10^{-6}$ for example. However in geodynamics velocities are of the order of a $\mathrm{cm\,yr}^{-1}$ which is about $3.1 \cdot 10^{-10} \mathrm{m\,s}^{-1}$. Small velocity changes would then be enforced only if $tol < 10^{-12}$. This value might prove completely unpractical for other applications. In light of all this one then resorts to assessing the *relative* change in the velocity by normalising the previous quantity by the average velocity in the domain $||\vec{\mathcal{V}}^i||$.

This is the very approach taken by Spiegelman *et al.* [1267] who monitor the relative changes in the solution from iteration to iteration:

$$\frac{||\Delta\vec{\mathcal{V}}||_{L2}}{||\vec{\mathcal{V}}||_{L2}} = \left(\frac{\int\limits_{\Omega}(\vec{\mathcal{V}}_i - \vec{\mathcal{V}}_{i-1}) \cdot (\vec{\mathcal{V}}_i - \vec{\mathcal{V}}_{i-1})dV}{\int\limits_{\Omega}\vec{\mathcal{V}}_i \cdot \vec{\mathcal{V}}_i dV}\right)^{1/2}$$

Of course, if a mixed formulation is used where velocity and pressure are solved for unknowns, the same monitoring can be done for pressure:

$$\frac{||\Delta\vec{\mathcal{P}}||_{L2}}{||\vec{\mathcal{P}}||_{L2}} = \left(\frac{\int\limits_{\Omega}(\vec{\mathcal{P}}_i - \vec{\mathcal{P}}_{i-1}) \cdot (\vec{\mathcal{P}}_i - \vec{\mathcal{P}}_{i-1})dV}{\int\limits_{\Omega}\vec{\mathcal{P}}_i \cdot \vec{\mathcal{P}}_i dV}\right)^{1/2}$$

Convergence is reached when both are below 0.001 (as in Lemiale *et al.* (2008) [810]) or 0.0001 (as in Kaus *et al.* (2010) [720]).

The last option is via the nonlinear residual. Coming back to the penalty formulation, we can form the nonlinear residual as follows:

$$\vec{\mathcal{R}}^i = \mathbb{K}(\eta_{\text{eff}}(\dot{\vec{\varepsilon}}^i, p^i)) \cdot \vec{\mathcal{V}}^i - \vec{f}$$

where $\mathbb{K}$ is defined in Section 10.7. Close to convergence $\vec{\mathcal{V}}^i$ and $\vec{\mathcal{V}}^{i-1}$ are very close so that we expect the residual $\vec{\mathcal{R}}$ to become smaller and smaller. In order to extract a scalar from $\vec{\mathcal{R}}$ we once again resort to the L^2-norm and we also wish to to monitor its relative change. In this case it is customary to use $\vec{\mathcal{R}}^0 = \vec{f}$ so that the convergence criterion becomes

$$\frac{||\vec{\mathcal{R}}^i||}{||\vec{\mathcal{R}}^0||} < tol.$$

FINISH: explain problem with mixed formulation!

---

[6]Two caveats here: the amplitude of the chequerboard mode might come into play - in the case $Q_1 \times P_0$ elements are used- and so does the applied smoothing.
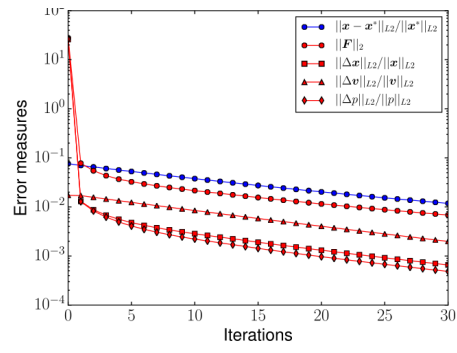
Taken from Spiegelman, May & Wilson *et al.* (2016). Example of reported nonlinear convergence.

16.4 Strain weakening

Several mechanisms may contribute to strain or strain rate dependent weakening but their relative and absolute importance is poorly constrained. Furthermore, weakening mechanisms are often crudely parameterised in geodynamical codes with simple mathematical functions and a limited number of parameters.

For example, Allken, Huismans, and Thieulot [10] (2011) authors use a von Mises plasticity formulation so that the rheology is parameterised by the cohesion c, or $c = \sigma_y$ in their notations. The yield strength σ_y starts is constant until the strain ε reaches the threshold value ε_1. It then decreases linearly from σ_y to σ_y^{sw} between ε_1 and ε_2. For strain values $\varepsilon > \varepsilon_2$, the yield strength remains constant at σ_y^{sw} .
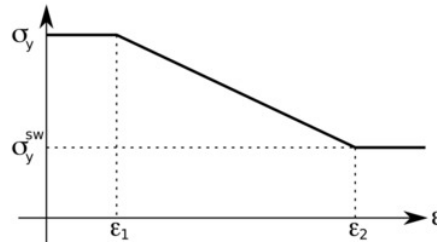
Taken from Allken, Huismans, and Thieulot [10] (2011).

The same authors in a subsequent study use a Drucker-Prager rheology parameterised by cohesion c and friction angle ϕ. They use the same approach as before but now both parameters are subjected to strain weakening:
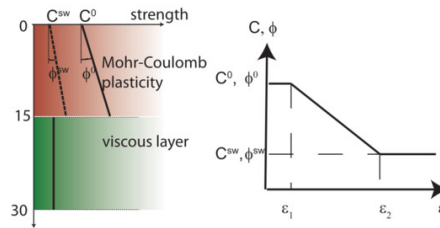
Taken from Allken, Huismans, and Thieulot [9] (2012), see also Thieulot [1343] (2011).

They further define the factor $R = C^0/C^{sw} = \phi^0/\phi^{sw} \geq 1$ which is a proxy for the ratio σ_y/σ_y^{sw} where $\sigma_y = p \sin \phi + c \, \cos \phi$, and carry out 3D crustal extensional models for R between 2 and 5.

- In Le Pourhiet, May, Huille, Watremez, and Leroy [801] (2017) the authors also define

$$\tau_y = p \sin(\phi(\varepsilon^p)) + c_0 \cos(\phi(\varepsilon^p))$$

but the cohesion is regarded to be constant. The angle of friction ϕ is assumed to decrease as a function of the accumulated plastic strain ε^p to

$$\phi(\varepsilon^p) = \max \left(\phi_\infty, \phi_0 - \frac{\varepsilon^p(\phi_0 - \phi_\infty)}{\varepsilon_\infty^p} \right)$$

This equation defines an empirical softening relation which reduces the friction angle linearly with accumulated plastic strain. ϕ_0 defines the initial friction angle, ε_∞^p represents the measure of plastic strain after which complete softening is achieved and internal friction angle reaches ϕ_∞ . Plastic strain represents an integrated, tensorial invariant measure of the deformation which has occurred due to plastic yielding. Thus, the quantity ε^p can be regarded as a simplified measure of material damage.

693

- In Dyksterhuis *et al.* Dyksterhuis, Rey, Mueller, and Moresi [379] (2007) a variant of the above formulation is used:

$$f(\varepsilon) = \begin{cases} 1 - (1-a)(\varepsilon/\varepsilon_0)^n & \varepsilon \leq \varepsilon_0 \\ a & \varepsilon \geq \varepsilon_0 \end{cases}$$

where ε is the accumulated plastic strain, ε_0 is the saturation strain beyond which no further weakening takes place, n is an exponent that controls the shape of the function and a is a maximum value of strain weakening beyond which no further weakening occurs. This equation leads to the following plot:
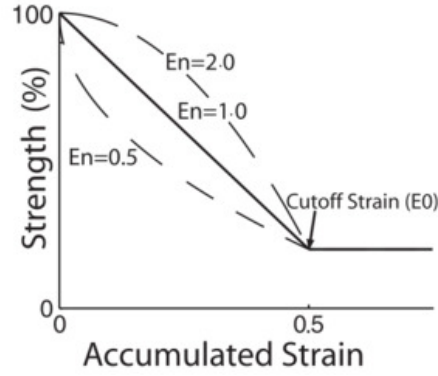
Strain-softening behaviour showing strength weakening from 100 to 20% after an accumulated strain of 0.5, after which no further weakening occurs.

Dashed lines show the effect of the exponential parameter (En) on the curve. Taken from Dyksterhuis, Rey, Mueller, and Moresi [379] (2007).

Although it is not specified in Dyksterhuis, Rey, Mueller, and Moresi [379] what f is, other users of the code specify that the yield strength is given by

$$\sigma_y = (B_0 + B_1 p) f(\varepsilon)$$

where p is the pressure, B_0 is the cohesion, or yield stress at zero pressure, and B_p is the pressure dependence of the yield stress, equivalent to the friction coefficient in Byerlee's law.

In Yang, Moresi, Zhao, Sandiford, and Whittaker [1481] (2018) the authors take a different approach:

$$C = C_0 + C_1 \exp\left(-\frac{\varepsilon_{plast}}{\varepsilon_{ref}}\right)$$

$$\mu = \mu_0 + \mu_1 \exp\left(-\frac{\varepsilon_{plast}}{\varepsilon_{ref}}\right)$$

where C_0 and $C_0 + C_1$ represent the minimum and maximum cohesions, respectively; μ_0 and $\mu_0 + \mu_1$ represent the minimum and maximum frictional coefficients, respectively. ε_{plast} and ε_{ref} represent accumulated plastic strain and reference strain, respectively.

- In Leroy and Ortiz [818] (1989) the authors describe another formulation for plastic hardening. The angle of friction changes with the accumulated plastic strain:

$$\sin\phi = \sin\phi_i + \frac{2(\sin\phi_f - \sin\phi_i)\sqrt{\varepsilon_c^p \varepsilon^p}}{\varepsilon^p + \varepsilon_c^p}$$

where ϕ transitions from an initial value ϕ_i to a maximum ϕ_f attained when the effective plastic strain reaches a critical value ε_c^p. When $\varepsilon^p \to \varepsilon_c^p$ then $\phi \to \phi_f$.

Relevant Literature: Sterpi [1295] (1999), Nijholt and Govers [999] (2015).

16.5 Assigning values to quadrature points

As we have seen in Section 10, the building of the elemental matrix and rhs requires (at least) to assign a density and viscosity value to each quadrature point inside the element. Depending on the type of modelling, this task can prove more complex than one might expect and have large consequences on the solution accuracy.

Here are several options:

- The simplest way (which is often used for benchmarks) consists in computing the 'real' coordinates (x_q, y_q, z_q) of a given quadrature point based on its reduced coordinates (r_q, s_q, t_q), and passing these coordinates to a function which returns density and/or viscosity at this location. For instance, for the Stokes sphere:

```
def rho(x,y):
    if (x-.5)**2+(y-0.5)**2<0.123**2:
        val=2.
    else:
        val=1.
    return val

def mu(x,y):
    if (x-.5)**2+(y-0.5)**2<0.123**2:
        val=1.e2
    else:
        val=1.
    return val
```

  This is very simple, but it has been shown to potentially be problematic. In essence, it can introduce very large contrasts inside a single element and perturb the quadrature. Please read section 3.3 of [593] and/or have a look at the section titled "Averaging material properties" in the ASPECT manual.

- another similar approach consists in assigning a density and viscosity value to the nodes of the FE mesh first, and then using these nodal values to assign values to the quadrature points. Very often ,and quite logically, the basis functions are used to this effect. Indeed we have seen before that for any point (r, s, t) inside an element we have

$$f_h(r, s, t) = \sum_i^m f_i N_i(r, s, t)$$

  where the f_i are the nodal values and the N_i the corresponding basis functions.

  In the case of linear elements (Q_1 basis functions), this is straightforward. In fact, the basis functions N_i can be seen as moving weights: the closer the point is to a node, the higher the weight (basis function value).

  However, this is quite another story for quadratic elements (Q_2 basis functions). In order to illustrate the problem, let us consider a 1D problem. The basis functions are

$$N_1(r) = \frac{1}{2}r(r - 1) \qquad N_2(r) = 1 - r^2 \qquad N_3(r) = \frac{1}{2}r(r + 1)$$

  Let us further assign: $\rho_1 = \rho_2 = 0$ and $\rho_3 = 1$. Then

$$\rho_h(r) = \sum_i^m \rho_i N_i(r) = N_3(r)$$

There lies the core of the problem: the $N_3(r)$ basis function is negative for $r \in [-1, 0]$. This means that the quadrature point in this interval will be assigned a negative density, which is nonsensical and numerically problematic!

use 2X Q1. write about it !

The above methods work fine as long as the domain contains a single material. As soon as there are multiple fluids in the domain a special technique is needed to track either the fluids themselves or their interfaces. Let us start with markers. We are then confronted to the infernal trio (a *menage a trois*?) which is present for each element, composed of its nodes, its markers and its quadrature points.

Each marker carries the material information (density and viscosity). This information must ultimately be projected onto the quadrature points. Two main options are possible: an algorithm is designed and projects the marker-based fields onto the quadrature points directly or the marker fields are first projected onto the FE nodes and then onto the quadrature points using the techniques above.

————————————

At a given time, every element e contains n^e markers. During the FE matrix building process, viscosity and density values are needed at the quadrature points. One therefore needs to project the values carried by the markers at these locations. Several approaches are currently in use in the community and the topic has been investigated by [350] and [375] for instance.

ELEFANT adopts a simple approach: viscosity and density are considered to be elemental values, i.e. all the markers within a given element contribute to assign a unique constant density and viscosity value to the element by means of an averaging scheme.

While it is common in the literature to treat the so-called arithmetic, geometric and harmonic means as separate averagings, I hereby wish to introduce the notion of generalised mean, which is a family of functions for aggregating sets of numbers that include as special cases the arithmetic, geometric and harmonic means.

If p is a non-zero real number, we can define the generalised mean (or power mean) with exponent p of the positive real numbers $a_1, \ldots a_n$ as:

$$M_p(a_1, \ldots a_n) = \left(\frac{1}{n} \sum_{i=1}^{n} a_i^p \right)^{1/p} \tag{16.14}$$

and it is trivial to verify that we then have the special cases:

$$M_{-\infty} = \lim_{p \to -\infty} M_p = \min(a_1, \ldots a_n) \quad \text{(minimum)} \tag{16.15}$$

$$M_{-1} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}} \quad \text{(harm. avrg.)} \tag{16.16}$$

$$M_0 = \lim_{p \to 0} M_p = \left(\prod_{i=1}^{n} a_i \right)^{1/n} \quad \text{(geom. avrg.)} \tag{16.17}$$

$$M_{+1} = \frac{1}{n} \sum_{i=1}^{n} a_i \quad \text{(arithm. avrg.)} \tag{16.18}$$

$$M_{+2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} a_i^2} \quad \text{(root mean square)} \tag{16.19}$$

$$M_{+\infty} = \lim_{p \to +\infty} M_p = \max(a_1, \ldots a_n) \quad \text{(maximum)} \tag{16.20}$$

Note that the proofs of the limit convergence are given in [183].

An interesting property of the generalised mean is as follows: for two real values p and q, if $p < q$ then $M_p \leq M_q$. This property has for instance been illustrated in Fig. 20 of [1200].

One can then for instance look at the generalised mean of a randomly generated set of 1000 viscosity values within $10^{18}Pa.s$ and $10^{23}Pa.s$ for $-5 \leq p \leq 5$. Results are shown in the figure hereunder and the arithmetic, geometric and harmonic values are indicated too. The function M_p assumes an arctangent-like shape: very low values of p will ultimately yield the minimum viscosity in the array while very high values will yield its maximum. In between, the transition is smooth and occurs essentially for $|p| \leq 5$.
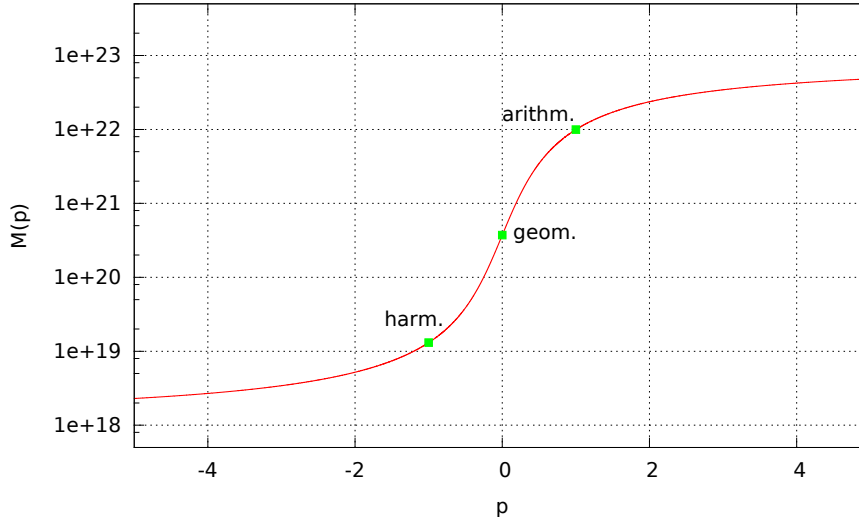
▷ *python_codes/fieldstone_markers_avrg*

16.6 Matrix (Sparse) storage

The FE matrix[7] (or the blocks which compose it) is the result of the assembly process of all elemental matrices. Its size can become quite large when the resolution is being increased (from thousands of lines/columns to tens of millions, if not even more).

One important property of the matrix is its sparsity. Typically musc *much* less than 1% of the matrix terms is not zero and this means that the matrix storage can and *should* be optimised. Clever storage formats were designed early on since the amount of RAM memory in computers was the limiting factor 3 or 4 decades ago [1166].

There are several standard formats[8], e.g.:

- compressed sparse row format (CSR)

- compressed sparse column format (CSC)

- the Coordinate Format (COO)

- Skyline Storage Format

I focus on the CSR format in what follows since it is the most common format and it is the one used in ELEFANT and in almost all code of FieldStone. In the following sections I look at typical cases in 2d and 3d.

16.6.1 2d domain - Q_1 - One degree of freedom per node

Let us consider again the 3×2 element grid which counts 12 nodes.

$nelx = 3, nnx = 4$
$nely = 2, nny = 3$
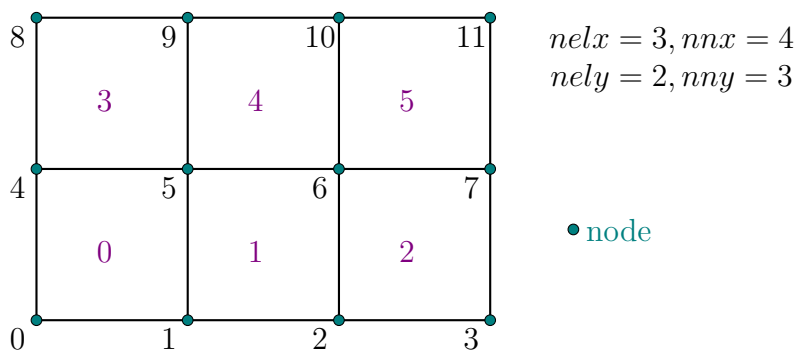
In the case there is only a single degree of freedom per node, the assembled FEM matrix $\boldsymbol{A}$ will look

---

[7] what follows also applied to matrices obtained by means of the FD method
[8] https://en.wikipedia.org/wiki/Sparse_matrix

like this:

$$
\boldsymbol{A} =
\begin{pmatrix}
\square & \square & & & \square & \square & & & & & & \\
\square & \square & \square & & \square & \square & \square & & & & & \\
& \square & \square & \square & & \square & \square & \square & & & & \\
& & \square & \square & & & \square & \square & & & & \\
\square & & & & \square & \square & & & \square & \square & & \\
\square & \square & & & \square & \square & \square & & \square & \square & \square & \\
& \square & \square & \square & & \square & \square & \square & & \square & \square & \square \\
& & \square & \square & & & \square & \square & & & \square & \square \\
& & & & \square & \square & & & \square & \square & & \\
& & & & \square & \square & \square & & \square & \square & \square & \\
& & & & & \square & \square & \square & & \square & \square & \square \\
& & & & & & \square & \square & & & \square & \square \\
\end{pmatrix}
$$

where the $\square$ stand for non-zero terms. This matrix structure is actually very predictable and stems from the fact that

- node 0 is connected to nodes 0,1,4,5 (1st line of the matrix)

- node 1 is connected to nodes 0,1,2,4,5,6 (2nd line of the matrix)

- node 2 is connected to nodes 1,2,3,5,6,7 (3rd line of the matrix)

- node 3 is connected to nodes 2,3,6,7 (4th line of the matrix)

- node 4 is connected to nodes 0,1,4,5,8,9 (5th line of the matrix)

- node 5 is connected to nodes 0,1,2,4,5,6,8,9,10 (6th line of the matrix)

- node 6 is connected to nodes 1,2,3,5,6,7,9,10,11 (7th line of the matrix)

- node 7 is connected to nodes 2,3,6,7,10,11 (8th line of the matrix)

- node 8 is connected to nodes 4,5,8,9 (9th line of the matrix)

- node 9 is connected to nodes 4,5,6,8,9,10 (10th line of the matrix)

- node 10 is connected to nodes 5,6,7,9,10,11 (11th line of the matrix)

- node 11 is connected to nodes 6,7,10,11 (last line/column of the matrix)

By 'connected' we mean that there is In light thereof, we have

- 4 corner nodes which have 4 neighbours (counting themselves),

- $2(nnx - 2)$ nodes which have 6 neighbours,

- $2(nny - 2)$ nodes which have 6 neighbours,

- $(nnx - 2) \times (nny - 2)$ nodes which have 9 neighbours.

In total, the number of non-zero terms in the matrix above is then:

$$
NZ = 4 \times 4 + 4 \times 6 + 2 \times 6 + 2 \times 9 = 70
$$

and in general, we would then have:

$$
NZ = 4 \times 4 + [2(nnx - 2) + 2(nny - 2)] \times 6 + (nnx - 2)(nny - 2) \times 9
$$

Let us temporarily assume $nnx = nny = n$. The matrix size (total number of unknowns) is then $N = n^2$ and
$$NZ = 16 + 24(n-2) + 9(n-2)^2$$
A full matrix array would contain $N^2 = n^4$ terms. The ratio of NZ (the actual number of reals to store) to the full matrix size (the number of reals a full matrix contains) is then
$$R = \frac{16 + 24(n-2) + 9(n-2)^2}{n^4}$$
It is then obvious that when n is large enough $R \sim 1/n^2$.

CSR stores the nonzeros of the matrix row by row, in a single indexed array A of double precision numbers. Another array COLIND (also often called ja) contains the column index of each corresponding entry in the A array. A third integer array RWPTR (also often called ia) contains pointers to the beginning of each row, which an additional pointer to the first index following the nonzeros of the matrix A. A and COLIND have length NZ and RWPTR has length N+1.

In the case of the here-above matrix, the arrays COLIND and RWPTR will look like:

$$\mathsf{ja=COLIND} \quad = \quad (0, 1, 4, 5, \; 0, 1, 2, 4, 5, 6, \; 1, 2, 3, 5, 6, 7, ..., 6, 7, 10, 11)$$
$$\mathsf{ia=RWPTR} \quad = \quad (0, 4, 10, 16, ...)$$

16.6.2   2d domain - Q_1 - Symmetric matrix CSR storage

If the matrix is symmetric, i.e. $\boldsymbol{A} = \boldsymbol{A}^T$, then we may wish to only store half of it, always in the interest of saving memory. Only the following remaining $\square$ entries are relevant now:

$$\boldsymbol{A} = \begin{pmatrix} \square & \square & & & \square & \square & & & & & & \\ & \square & \square & & \square & \square & \square & & & & & \\ & & \square & \square & & \square & \square & \square & & & & \\ & & & \square & & & \square & \square & & & & \\ & & & & \square & \square & & & \square & \square & & \\ & & & & & \square & \square & & \square & \square & \square & \\ & & & & & & \square & \square & & \square & \square & \square \\ & & & & & & & \square & & & \square & \square \\ & & & & & & & & \square & \square & & \\ & & & & & & & & & \square & \square & \\ & & & & & & & & & & \square & \square \\ & & & & & & & & & & & \square \end{pmatrix}$$

We see that the number of nonzeros is now
$$NZ_{symm} = \frac{NZ - N}{2} + N$$
and in this case $NZ_{symm} = (70 - 12)/2 + 12 = 41$. Then

$$\mathsf{ja=COLIND} \quad = \quad (0, 1, 4, 5, \; 1, 2, 4, 5, 6, \; 3, 5, 6, 7, ..., , 11)$$
$$\mathsf{ia=RWPTR} \quad = \quad (0, 4, 9, 14, ...)$$

In case the numbering is Fortran-like, then

$$\mathsf{ja=COLIND} \quad = \quad (1, 2, 5, 6, 2, 3, 5, 6, 7, 3, 4, 6, 7, 8, 4, 7, 8, 5, 6, 9, 10, 6, 7, 9, 10, 11,$$
$$7, 8, 10, 11, 12, 8, 11, 12, 9, 10, 10, 11, 11, 12, 12)$$
$$\mathsf{ia=RWPTR} \quad = \quad (1, 5, 10, 15, 18, 22, 27, 32, 35, 37, 39, 41, 42)$$

16.6.3 2d domain - Q_1 - Two degrees of freedom per node

When there are two degrees of freedom per node (i.e. $ndofV = 2$), such as in the case of the Stokes equation in two-dimensions, the size of the $\mathbb{K}$ matrix is given by $NfemV = nnx * nny * ndofV$ where $NfemV$ is the total number of velocity degrees of freedom.
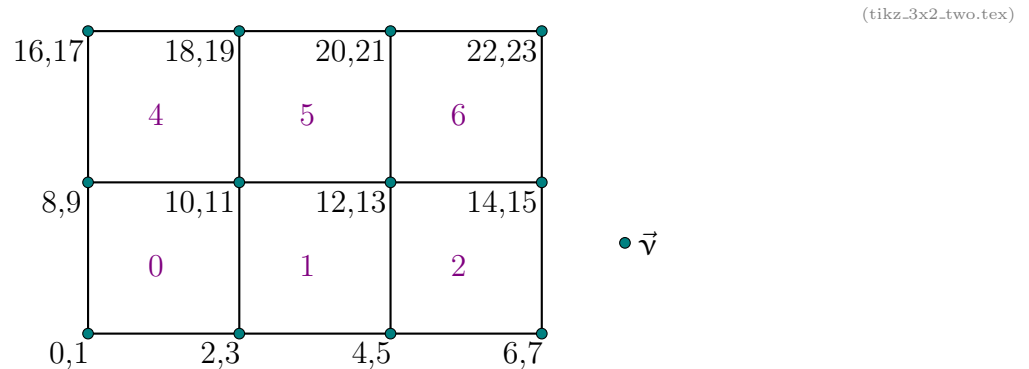
In the case of the small grid above, we have then $NfemV = 24$ and elemental matrices are now 8×8 in size.

We still have

- 4 corner nodes which have 4 neighbours

- $2(nnx - 2)$ nodes which have 6 neighbours

- $2(nny - 2)$ nodes which have 6 neighbours

- $(nnx - 2) \cdot (nny - 2)$ nodes which have 9 neighbours,

but now each degree of freedom from a node sees the other two degrees of freedom of another node too. In that case, the number of nonzeros has been multiplied by four and the assembled FEM

matrix looks like:

Note that the degrees of freedom are organised as follows:

$$(u_0, v_0, u_1, v_1, u_2, v_2, ...u_{11}, v_{11})$$

In general, we would then have:

$$NZ = 4\left[4 \times 4 + [2(nnx - 2) + 2(nny - 2)] \times 6 + (nnx - 2)(nny - 2) \times 9\right]$$

and in the case of the small grid, the number of non-zero terms in the matrix is then:

$$NZ = 4\left[4 \times 4 + 4 \times 6 + 2 \times 6 + 2 \times 9\right] = 280$$

In the case of the here-above matrix, the arrays COLIND and RWPTR will look like:

$$
\begin{aligned}
\mathsf{ja=COLIND} &= (0, 1, 2, 3, 8, 9, 10, 11, \ 0, 1, 2, 3, 8, 9, 10, 11, \ ...) \\
\mathsf{ia=RWPTR} &= (0, 8, 16, 28, ...)
\end{aligned}
$$

Assuming we are using $Q_1 \times P_0$ elements, the structure of the matrix $\mathbb{G}_{el}^T$ is as follows (the 6 pressure dofs are connected to 24 velocity dofs):

(16.21)

702
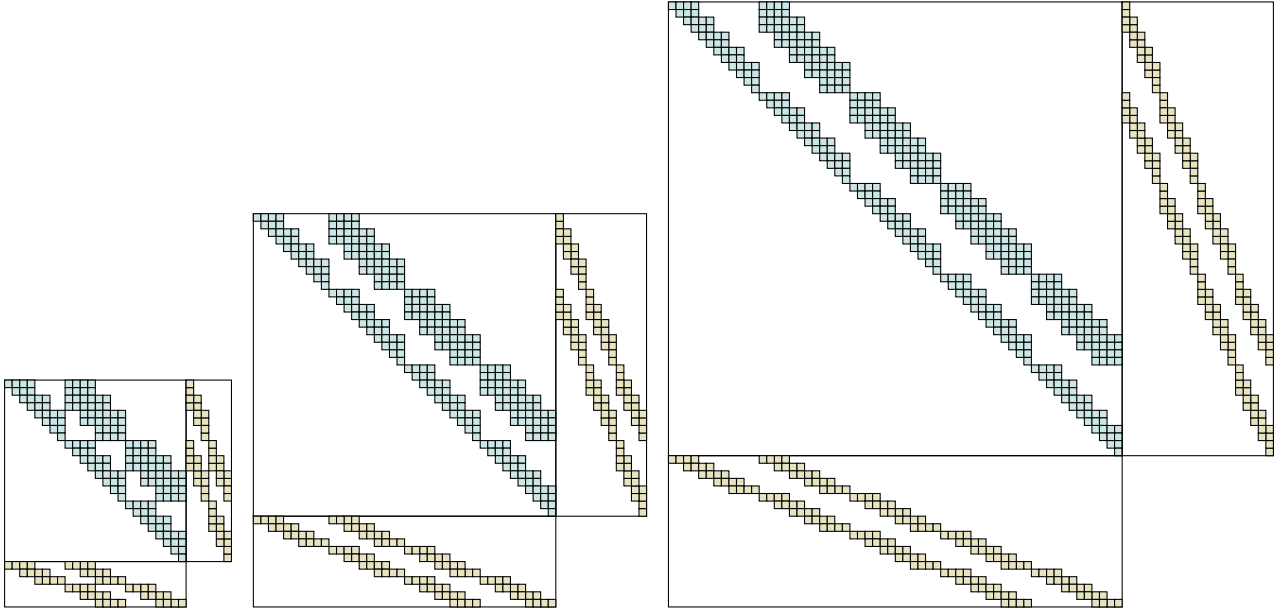
From left to right: Nonzero structures of the assembled Stokes matrix for a 3×2, 4×3 and 5×4 mesh of $Q_1 \times P_0$ elements.

Assuming we are now using $Q_1 \times Q_1$ elements (without bubble[9]), the structure of the matrix $\mathbb{G}_{el}^T$ is different: we now have 12 pressure dofs which are coupled to 24 velocity dofs:

$$
\begin{pmatrix}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\
0 & \square & \square & \square & \square & & & & & \square & \square & \square & \square & & & & & & & & & & & & \\
1 & \square & \square & \square & \square & \square & \square & & & \square & \square & \square & \square & \square & \square & & & & & & & & & & \\
2 & & & \square & \square & \square & \square & \square & \square & & & \square & \square & \square & \square & \square & \square & & & & & & & & \\
3 & & & & & \square & \square & \square & \square & & & & & \square & \square & \square & \square & & & & & & & & \\
... & & & & & & & & & & & & & & & & & & & & & & & & \\
9 & & & & & & & & & \square & \square & \square & \square & \square & \square & & & & & \square & \square & \square & \square & & \\
10 & & & & & & & & & & & \square & \square & \square & \square & \square & \square & & & \square & \square & \square & \square & \square & \square \\
11 & & & & & & & & & & & & & \square & \square & \square & \square & & & & & \square & \square & \square & \square \\
\end{pmatrix}
\tag{16.22}
$$

If now the velocity dofs are organised as follows

$$\left(u_0, u_1, u_2, ..., u_{11}, v_0, v_1, v_2, ..., v_{11} \right)$$

then the sparsity pattern of the assembled $\mathbb{K}$ matrix looks like

---

[9]This element pair is obviously unstable and should not be used.

$$
\begin{pmatrix}
\begin{array}{cccccccccccccccccccccccc}
u & u & u & u & u & u & u & u & u & u & u & u & v & v & v & v & v & v & v & v & v & v & v & v \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12
\end{array}
\end{pmatrix}
$$

(16.23)

We see that in this case the matrix is composed of 4 identical blocks, the same blocks we have previously obtained when each node only had a single dof. In other words,

$$
\mathbb{K} = \begin{pmatrix} \mathbb{K}_{xx} & \mathbb{K}_{xy} \\ \mathbb{K}_{yx} & \mathbb{K}_{yy} \end{pmatrix} = \begin{pmatrix} \mathbb{K}_1 & \mathbb{K}_1 \\ \mathbb{K}_1 & \mathbb{K}_1 \end{pmatrix}
$$

16.6.4   2d domain - Q_2 - Two degrees of freedom per node

When there are now two degrees of freedom per node, such as in the case of the Stokes equation in two-dimensions, the size of the $\mathbb{K}$ matrix is given $NfemV = nnx * nny * ndofV$ where $NfemV$ is the total number of velocity degrees of freedom. What is different here is that for Q_2 elements we have $nnx = 2 * nelx + 1$ and $nny = 2 * nely + 1$.

704

In the case of the small grid above, we have then $nelx = 3$, $nely = 2$, so that $nnx = 7$ and $nny = 5$, and then $NfemV = 7*5*2 = 70$ and elemental matrices are now 18×18 in size.

Let us start with the simple case of 1 dof per node:
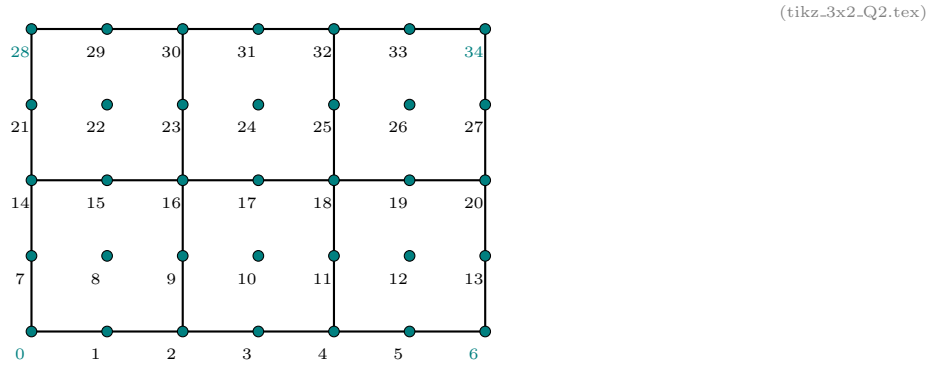
Concretely here:

- nodes 0,6,28,34 see 9 nodes (corners)

- nodes 1,3,5,7,8,10,12,13,21,22,24,26,27,29,31,33 see 9 nodes

- nodes 2,4,9,11,14,15,17,19,20,23,25,30,32, see 15 nodes

- nodes 16,18 see 25 nodes

If there was only one dof per node, we would find the number of non zeros as follow:

$$NZ = 4*9 + 16*9 + 13*15 + 2*25 = 36 + 144 + 195 + 50 = 425$$

But since there are two velocity dofs per node, we find that the total number of nonzeros is 4 times higher, i.e.

$$NZ = 1700$$

And if we choose for a symmetric CSR storage:

$$NZ_{symm} = \frac{NZ - N}{2} + N = \frac{1700 - 70}{2} + 70 = 885$$

Let us now turn to the real case of 2 dofs per node and establish who 'sees' who:

dof	sees other dofs	total
0	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
1	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
2	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
3	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
4	0,1,2,3,4,5,6,7,8,9,14,15,16,17,18,19,20,21,22,23,28,29,30,31,32,33,34,35,36,37	30
5	0,1,2,3,4,5,6,7,8,9,14,15,16,17,18,19,20,21,22,23,28,29,30,31,32,33,34,35,36,37	30
6	4,5,6,7,8,9,18,19,20,21,22,23,32,33,34,35,36,37	18
7	4,5,6,7,8,9,18,19,20,21,22,23,32,33,34,35,36,37	18
8	4,5,6,7,8,9,10,11,12,13,18,19,20,21,22,23,24,25,26,27,32,33,34,35,36,37,38,39,40,41	30
9	4,5,6,7,8,9,10,11,12,13,18,19,20,21,22,23,24,25,26,27,32,33,34,35,36,37,38,39,40,41	30
10	8,9,10,11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	18
11	8,9,10,11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	18
12	8,9,10,11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	18
13	8,9,10,11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	18
14	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
15	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
16	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
17	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
...	...	...
68	36,37,38,39,50,51,52,53,54,55,64,65,66,67,68,69	18
69	36,37,38,39,50,51,52,53,54,55,64,65,66,67,68,69	18

The second column of this array is the content of the `ja` array if a full storage is used. If we now use a symmetric storage then:

dof	sees other dofs	total
0	0,1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	18
1	1,2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	17
2	2,3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	16
3	3,4,5,14,15,16,17,18,19,28,29,30,31,32,33	15
4	4,5,6,7,8,9,14,15,16,17,18,19,20,21,22,23,28,29,30,31,32,33,34,35,36,37	
5	5,6,7,8,9,14,15,16,17,18,19,20,21,22,23,28,29,30,31,32,33,34,35,36,37	
6	6,7,8,9,18,19,20,21,22,23,32,33,34,35,36,37	
7	7,8,9,18,19,20,21,22,23,32,33,34,35,36,37	
8	8,9,10,11,12,13,18,19,20,21,22,23,24,25,26,27,32,33,34,35,36,37,38,39,40,41	
9	9,10,11,12,13,18,19,20,21,22,23,24,25,26,27,32,33,34,35,36,37,38,39,40,41	
10	10,11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	
11	11,12,13,22,23,24,25,26,27,36,37,38,39,40,41	
12	12,13,22,23,24,25,26,27,36,37,38,39,40,41	
13	13,22,23,24,25,26,27,36,37,38,39,40,41	
14	14,15,16,17,18,19,28,29,30,31,32,33	
15	15,16,17,18,19,28,29,30,31,32,33	
16	16,17,18,19,28,29,30,31,32,33	
17	17,18,19,28,29,30,31,32,33	
...	...	...
68	68,69	2
69	69	1

In order establish a pattern we will need a bigger mesh:
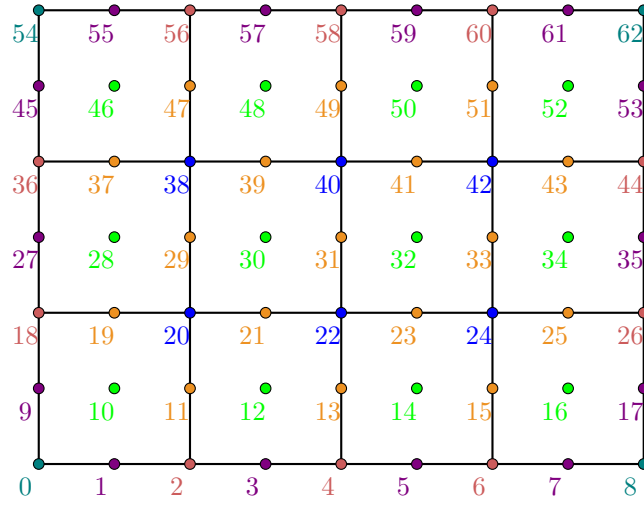
We have

- 4 corner nodes which have 9 neighbours

- nel mid-element nodes which have 9 neighbours

- $2 * nelx + 2 * nely$ mid-edge nodes on sides which have 9 neighbours

- $(nelx - 1) * nely + nelx * (nely - 1)$ internal mid-edges nodes which have 15 neighbours

- $2 * (nelx - 1) + 2 * (nely - 1)$ side nodes that have 15 neighbours

- $(nelx - 1) * (nely - 1)$ nodes which have 25 neighbours

In the end, the number of non-zeros (Q_1) is given by

$$
\begin{aligned}
NZ \;=\;& 4 * 9 \\
+\;& nel * 9 \\
+\;& (2 * nelx + 2 * nely) * 9 \\
+\;& [(nelx - 1) * nely + nelx * (nely - 1)] * 15 \\
+\;& [2 * (nelx - 1) + 2 * (nely - 1)] * 15 \\
+\;& (nelx - 1) * (nely - 1) * 25
\end{aligned}
$$

Verification: $nelx = 3$, $nely = 2$:

$$
\begin{aligned}
NZ \;=\;& 4 * 9 + 6 * 9 + (2 * 3 + 2 * 2) * 9 + [(3 - 1) * 2 + 3 * (2 - 1)] * 15 + [2 * (3 - 1) + 2 * (2 - 1)] * 15 + \\
=\;& 36 + 54 + 10 * 9 + 7 * 15 + 6 * 15 + 2 * 25 \\
=\;& 36 + 54 + 90 + 105 + 90 + 50 \\
=\;& 425
\end{aligned}
$$

as expected.

16.6.5   3d domain - Q_1 - CSR storage - One degree of freedom

Let us consider a $3 \times 4 \times 2$ grid which counts $nnx \cdot nny \cdot nnz = 5 \cdot 4 \cdot 3 = 60$ nodes. The assembled FEM matrix $\mathbb{K}$ size is then $N = nnx \times nny \times nnz \times ndof = 180$.
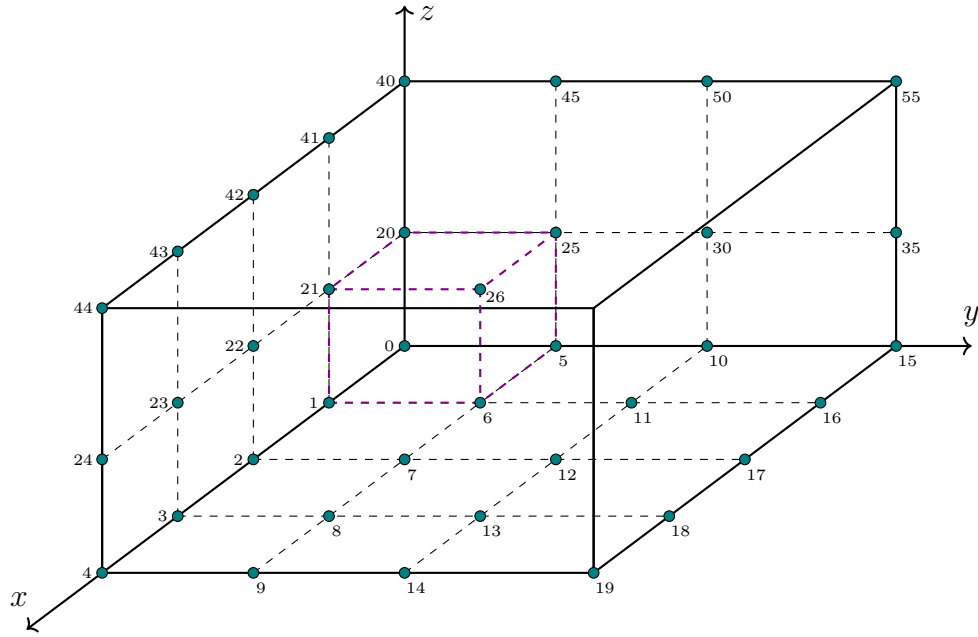
The total number of nonzeros in the case $ndof = 1$ would be decomposed as follows:

- 8 corners 'see' 8 neighbours

- 4 edges with $(nnx - 2)$ nodes in the x direction see 12 nodes

- 4 edges with $(nny - 2)$ nodes in the y direction see 12 nodes

- 4 edges with $(nnz - 2)$ nodes in the z direction see 12 nodes

- $2(nnx - 2)(nny - 2)$ nodes see 18 nodes

- $2(nnx - 2)(nnz - 2)$ nodes see 18 nodes

- $2(nny - 2)(nnz - 2)$ nodes see 18 nodes

- $(nnx - 2)(nny - 2)(nnz - 2)$ interior nodes see 27 nodes

16.6.6   3d domain - Q_2 - CSR storage - one degree of freedom
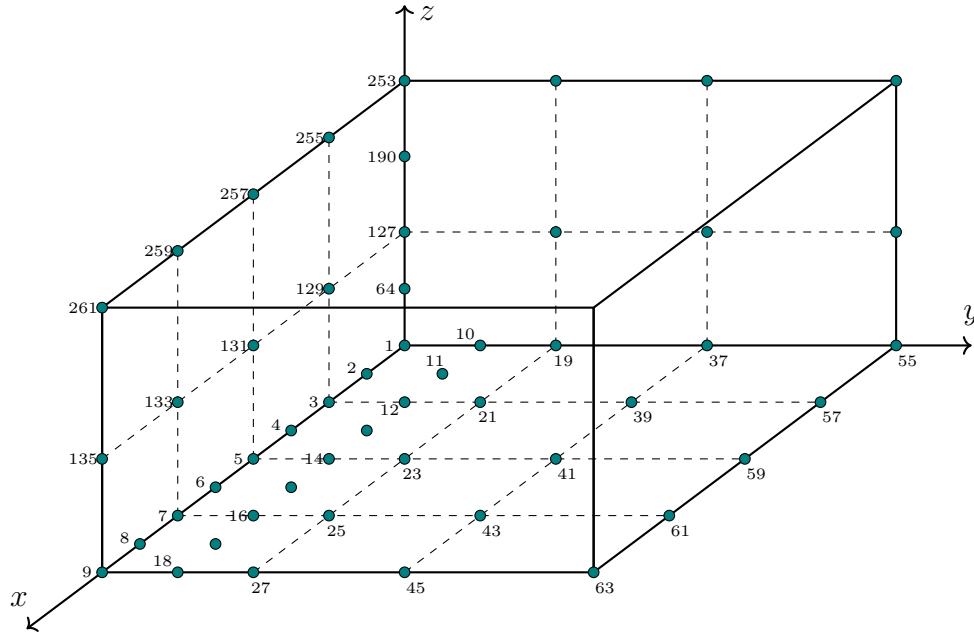
16.6.7   Matrix Storage in the Python codes of FieldStone

The majority of the early codes have the FE matrix being a full array

```
a_mat = np.zeros((Nfem,Nfem),dtype=np.float64)
```

and it is converted to CSR format on the fly in the solve phase:

```
sol = sps.linalg.spsolve(sps.csr_matrix(a_mat),rhs)
```

Note that linked list storages can be used (lil_matrix). Substantial memory savings but much longer compute times since it takes longer to write in such arrays. A conversion to CSR format is still necessary before calling the solver.

16.6.8   About Sparse Matrix-Vector multiplication

When the matrix A is stored in a two-dimensional array of size $N \times N$, its (left or right) multiplication by a vector is trivial. Either one resorts to writing a double for loop (not recommended), either one uses `numpy.dot`[10] in python, or `matmul` in Fortran.

However, when the matrix is stored as a single continuous array, say CSR, how does this work? This question is *very important* since iterative solvers such as the Conjugate Gradient solver (see Section 16.28) rely extensively on multiplying the matrix by many different vectors.

The Sparse Matrix-Vector multiplication operation is often abbreviated SpMV. To quote Knepley [755]: "The Sparse Matrix-Vector Product (SpMV) is today a workhorse of scientific computing. It is a central kernel in iterative linear and nonlinear solvers for PDE, and now for many graph algorithms." As explained in Williams *et al.* (2007) [1461] (and in many other sources on the topic), the algorithm for a basic SpMV implementation is rather simple in its naive form:

---

[10]`https://numpy.org/doc/stable/reference/generated/numpy.dot.html`

(a) algebra conceptualization   **(b)** CSR data structure   **(c)** CSR reference code
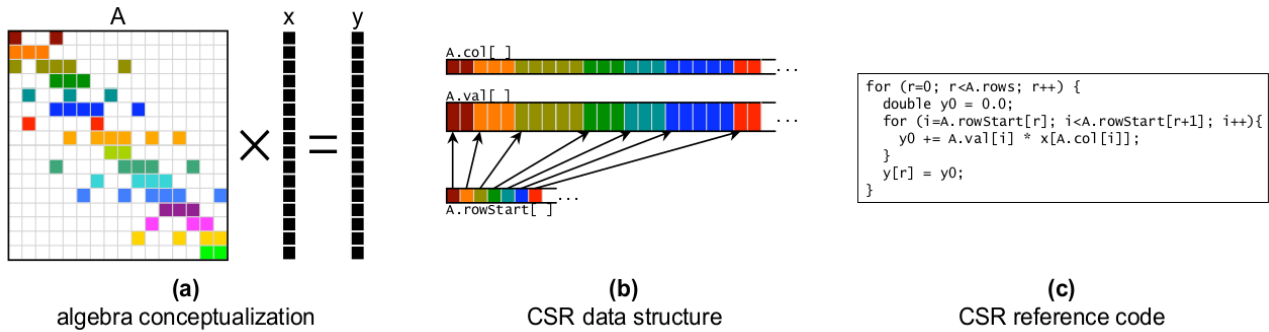
Taken from Williams *et al.* (2008) [1462]. Sparse Matrix Vector Multiplication (SpMV). (a) visualization of the algebra: $\vec{y} \leftarrow \boldsymbol{A} \cdot \vec{x}$. (b) Standard compressed sparse row (CSR) representation of the matrix. (c) The standard implementation of SpMV for a matrix stored in CSR. The outer loop is trivially parallelized without any data dependencies.

Let us assume that we wish to compute $\vec{y} = \boldsymbol{A} \cdot \vec{x}$ where $\boldsymbol{A}$ is in CSR format. The pseudo code then goes as follows:

```
for i in range(0,m):
    y0=0
    for k in range(ROWPTR[i],ROWPTR[i+1]):
        y0 += VAL[k] * x[COLIND[k]]
    y[i]=y0
```

Although technically correct, this algorithm is problematic because the vector x array is accessed indirectly and this causes a non-optimal use of the processor, which in the end makes the calculation take longer than it should.

The following piece of ForTran code comes from ELEFANT . Note that here (ROWPTR=ia, COLIND=ja, VAL=mat)

```
subroutine spmv (nr,nc,nz,x,y,mat,ja,ia)
implicit none
integer, intent(in)   :: nr,nc,nz
real(8), intent(in)   :: x(nc), mat(nz)
real(8), intent(out) :: y(nr)
integer, intent(in)   :: ja(nz),ia(nr+1)
real(8) t
integer i, k

do i = 1,nr
    t = 0.0d0
    do k=ia(i), ia(i+1)-1
        t = t + mat(k)*x(ja(k))
    end do
    y(i) = t
end do

end subroutine
```

How to make this calculation as efficiently as possible on CPUs and GPUs, on one thread or multiple threads has given rise to a lot of literature.

📚 Relevant Literature: Krotkiewski and Dabrowski [777] (2010), Section 9.4 of Knepley [755], Williams et al. [1462] (2008).

16.6.9   SpMV and SpMV-T with the CSR format - a concrete example

(What follows was orignally written for ELEFANT so that code excerpts and loop indexing are those of Fortran.)

Let us consider a simple matrix $\mathbb{G}$ which is not square (size is 3×5):

$$\mathbb{G}^T = \begin{pmatrix} 1 & 0 & 4 & 1 & 2 \\ 0 & 1 & 1 & 1 & 0 \\ 3 & 0 & 0 & 7 & 1 \end{pmatrix}$$

The number of rows is $nr = 3$, the number of columns is $nc = 5$ and the number of nonzeros is $nz = 10$.

Let us consider two vectors $\vec{\mathcal{V}}^T = (1, 1, 1, 1, 1)$ and $\vec{\mathcal{P}}^T = (1, 1, 1)$. Obviously, we have:

$$\mathbb{G}^T \cdot \vec{\mathcal{V}} = \begin{pmatrix} 8 \\ 3 \\ 11 \end{pmatrix} \quad \text{and} \quad \mathbb{G} \cdot \vec{\mathcal{P}} = \begin{pmatrix} 4 \\ 1 \\ 5 \\ 9 \\ 3 \end{pmatrix}$$

The CSR storage of $\mathbb{G}^T$ requires three arrays: ia (integer, size $nr + 1$), ja (integer, size nz) and mat (real, size nz). In the case of the small matrix above:

$$\begin{aligned} ia &= (1, 5, 8, 11) \\ ja &= (1, 3, 4, 5, 2, 3, 4, 1, 4, 5) \\ mat &= (1, 4, 1, 2, 1, 1, 1, 3, 7, 1) \end{aligned}$$

The sparse matrix vector multiplication kernel SpMV for $\vec{y} = \boldsymbol{A} \cdot \vec{x}$ has been explained above, and it is trivial to carry out this algorithm by hand and verify that the vector y is given by $y^T = (8, 3, 11)$.

Let us now turn to an interesting problem. Is it possible with the same arrays ia, ja, mat to compute the multiplication of the transpose of the matrix with a vector? The answer is of course positive and the code is given hereunder:

```
y=0.d0
do i = 1,nr
    do k=ia(i), ia(i+1)-1
        y(ja(k))=y(ja(k))+mat(k)*x(i)
    end do
end do
```

Let us take $i = 1$. The variable k then goes from 1 to 4. The inner loop does:

```
y(1)=y(1)+mat(1)*x(1)
y(3)=y(3)+mat(2)*x(1)
y(4)=y(4)+mat(3)*x(1)
y(5)=y(5)+mat(4)*x(1)
```

Let us take $i = 2$. The variable k then goes from 5 to 7. The inner loop does:

```
y(2)=y(2)+mat(5)*x(2)
y(3)=y(3)+mat(6)*x(2)
y(4)=y(4)+mat(7)*x(2)
```

Let us take $i = 3$. The variable k then goes from 8 to 10. The inner loop does:

```
y(1)=y(1)+mat(8)*x(3)
y(4)=y(4)+mat(9)*x(3)
y(5)=y(5)+mat(10)*x(3)
```

So in total, we have:

```
y(1)=mat(1)*x(1)+mat(8)*x(3)
y(2)=mat(5)*x(2)
y(3)=mat(2)*x(1)+mat(6)*x(2)
y(4)=mat(3)*x(1)+mat(7)*x(2)+mat(9)*x(3)
y(5)=mat(4)*x(1)+mat(10)*x(3)
```

which is indeed the result of the transposed of the matrix multiplied by a vector $\vec{x}$.

Let us consider a simple matrix $\mathbb{K}$ which is square (size is 5×5):

$$\mathbb{K} = \begin{pmatrix} 1 & 0 & 4 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 \\ 4 & 0 & 0 & 7 & 1 \\ 1 & 1 & 7 & 4 & 0 \\ 2 & 0 & 1 & 0 & 5 \end{pmatrix}$$

In this case , NZ=16.

$$\begin{aligned} ia &= (1, 5, 7, 10, 14, 17) \\ ja &= (1, 3, 4, 5, \quad 2, 4, \quad 1, 4, 5, \quad 1, 2, 3, 4, \quad 1, 3, 5) \\ mat &= (1, 4, 1, 2, 1, 1, 4, 7, 1, 1, 1, 7, 4, 2, 1, 5) \end{aligned}$$

The sparse matrix vector multiplication kernel SpMV for $\vec{y} = \boldsymbol{A} \cdot \vec{x}$ is given as follows in its simplest form. Since the matrix is symmetric, there is no use to store the whole matrix. Its upper half (for instance) will do. In this case, NZ= and then

$$\begin{aligned} ia &= (1, 5, 7, 9, 10, 11) \\ ja &= (1, 3, 4, 5, \quad 2, 4, \quad 4, 5, \quad 4, \quad 5) \\ mat &= (1, 4, 1, 2, \quad 1, 1, \quad 7, 1, \quad 4, \quad 5) \end{aligned}$$

All is good and well until one wishes to multiply the real matrix by a vector. The SpMV routines described above will not work since it will return the upper half of the matrix multiplied by the vector.

One can then write a decicated algorithm:

```
do i = 1,nr

   ! multiply the upper half by the vector

   do k=ia(i), ia(i+1)-1
      y(i) = y(i) + mat(k)*x(ja(k))
   end do

   ! multiply the transpose of matrix by vector
   ! but omit diagonal

   do k=ia(i), ia(i+1)-1
      if (i/=ja(k)) then
         y(ja(k))=y(ja(k))+mat(k)*x(i)
      end if
```

```
   end do

end do
```

Example:

```
y(1)
=y(1) + mat(1)*x(ja(1)) + mat(2)*x(ja(2)) + mat(3)*x(ja(3)) + mat(4)*x(ja(4))
=y(1) + mat(1)*x(1) + mat(2)*x(3) + mat(3)*x(4) + mat(4)*x(5)
```

etc ...

Finish?

16.7 Mesh generation

Before basis functions can be defined and PDEs can be discretised and solved we must first tesselate the domain with polygons, e.g. triangles and quadrilaterals in 2D, tetrahedra, prisms and hexahedra in 3D.

When the domain is itself simple (e.g. a rectangle, a sphere, ...) the mesh (or grid) can be (more or less) easily produced and the connectivity array filled with straightforward algorithms [1344]. However, real life applications can involve extremely complex geometries (e.g. a bridge, a human spine, a car chassis and body, etc ...) and dedicated algorithms/softwares must be used (see [1356, 445, 1475]).

We usually distinguish between two broad classes of grids: structured grids (with a regular connectivity) and unstructured grids (with an irregular connectivity).
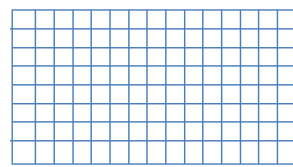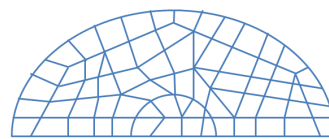
Structured Grid   Unstructured Grid

Remark. *Various families of so-called meshless methods exist and are commonly employed in Computational Fluid Dynamics [843, 829, 842, 844]. They are however very rarely used in Computational geodynamics, with a noticeable exception [564].*

16.7.1 Quadrilateral-based meshes

Let us now focus on the case of a rectangular computational domain of size Lx × Ly with a regular mesh composed of nelx×nely=nel quadrilaterals. There are then nnx×nny=nnp grid points. The elements are of size hx×hy with hx=Lx/nelx.

We have no reason to come up with an irregular/illogical node numbering so we can number nodes row by row or column by column as shown on the example hereunder of a 3×2 grid:

```
8=======9======10======11        2=======5=======8======11
|        |       |        |        |        |        |        |
|  (3)   |  (4)  |  (5)   |        |  (1)   |  (3)   |  (5)   |
|        |       |        |        |        |        |        |
4=======5======6=======7         1=======4=======7======10
|        |       |        |        |        |        |        |
|  (0)   |  (1)  |  (2)   |        |  (0)   |  (2)   |  (4)   |
|        |       |        |        |        |        |        |
0=======1======2=======3         0=======3=======6=======9


      "row by row"                  "column by column"
```

The numbering of the elements themselves could be done in a somewhat chaotic way but we follow the numbering of the nodes for simplicity. The row by row option is the adopted one in fieldstone and the coordinates of the points are computed as follows:

714

```
x = np.empty(nnp, dtype=np.float64)
y = np.empty(nnp, dtype=np.float64)
counter = 0
for j in range(0,nny):
    for i in range(0,nnx):
        x[counter]=i*hx
        y[counter]=j*hy
        counter += 1
```

The inner loop has `i` ranging from `0` to `nnx-1` first for `j=0, 1, ...` up to `nny-1` which indeed corresponds to the row by row numbering.

We now turn to the connectivity. As mentioned before, this is a structured mesh so that the so-called connectivity array, named `icon` in our case, can be filled easily. For each element we need to store the node identities of its vertices. Since there are `nel` elements and `m=4` corners, this is a `m×nel` array. The algorithm goes as follows:

```
icon =np.zeros((m,nel),dtype=np.int16)
counter = 0
for j in range(0,nely):
    for i in range(0,nelx):
        icon[0,counter] = i + j * nnx
        icon[1,counter] = i + 1 + j * nnx
        icon[2,counter] = i + 1 + (j + 1) * nnx
        icon[3,counter] = i + (j + 1) * nnx
        counter += 1
```

In the case of the 3×2 mesh, the `icon` is filled as follows:

element id→	0	1	2	3	4	5
node id↓						
0	0	1	2	4	5	6
1	1	2	3	5	6	7
2	5	6	7	9	10	11
3	4	5	6	8	9	10

It is to be understood as follows: element #4 is composed of nodes 5, 6, 10 and 9. Note that nodes are always stored in a counter clockwise manner, starting at the bottom left. This is very important since the corresponding basis functions and their derivatives will be labelled accordingly.

In three dimensions things are very similar. The mesh now counts `nelx×nely×nelz=nel` elements which represent a cuboid of size `Lx×Ly×Lz`. The position of the nodes is obtained as follows:

```
x = np.empty(nnp,dtype=np.float64)
y = np.empty(nnp,dtype=np.float64)
z = np.empty(nnp,dtype=np.float64)
counter=0
for i in range(0,nnx):
    for j in range(0,nny):
        for k in range(0,nnz):
            x[counter]=i*hx
            y[counter]=j*hy
            z[counter]=k*hz
            counter += 1
```

The connectivity array is now of size `m×nel` with `m=8`:

```
icon =np.zeros((m,nel),dtype=np.int16)
counter = 0
for i in range(0,nelx):
    for j in range(0,nely):
```

```
        for k in range(0,nelz):
            icon[0,counter]=nny*nnz*(i  )+nnz*(j  )+k
            icon[1,counter]=nny*nnz*(i+1)+nnz*(j  )+k
            icon[2,counter]=nny*nnz*(i+1)+nnz*(j+1)+k
            icon[3,counter]=nny*nnz*(i  )+nnz*(j+1)+k
            icon[4,counter]=nny*nnz*(i  )+nnz*(j  )+k+1
            icon[5,counter]=nny*nnz*(i+1)+nnz*(j  )+k+1
            icon[6,counter]=nny*nnz*(i+1)+nnz*(j+1)+k+1
            icon[7,counter]=nny*nnz*(i  )+nnz*(j+1)+k+1
            counter += 1
```

> produce drawing of node numbering

Although it is not very common in geosciences, quadrilateral meshes are sometimes employed in a boundary-fitted way, as shown hereunder:
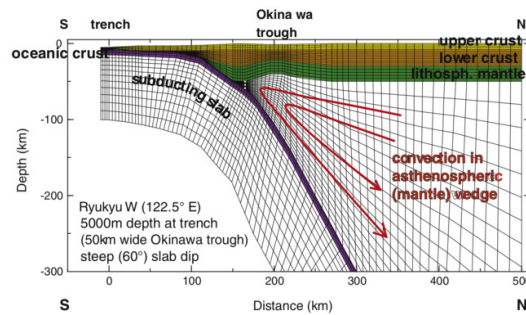
Taken from Gutscher et al. [547] (2016).

Relevant Literature: Joun and Lee [695] (1997)

16.7.2 Delaunay triangulation and Voronoi cells, and triangle-based meshes

The topic of Delaunay[11] triangulation is vast, but a simple definition can be written as follows: "a Delaunay triangulation for a set P of points in a plane is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P)." [wikipedia] Other properties of such triangulations are that they maximize the minimum angle of all the angles of the triangles in the triangulation. Note that for four or more points on the same circle (e.g., the vertices of a rectangle) the Delaunay triangulation is not unique and that points on a line also cannot yield a valid triangulation (for the simple reason that they do not form a triangle).
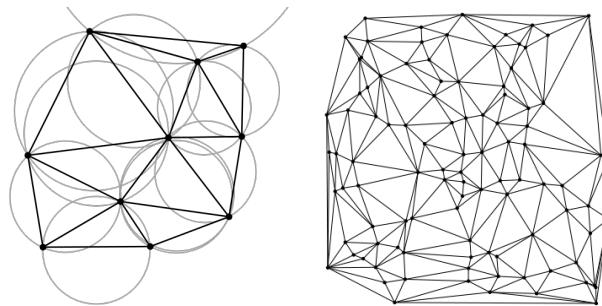
a) A Delaunay triangulation in the plane with circumcircles shown. b) The Delaunay triangulation of a random set of 100 points in a plane.

The Delaunay triangulation of a discrete point set P in general corresponds to the dual graph of the Voronoi diagram for P. A Voronoi diagram is composed of non-overlapping Voronoi cells which make a partition of the plane. For each point there is a corresponding region consisting of all points closer to that point than to any other: this region is the Voronoi cell of that point.

---

[11]The triangulation is named after Boris Delaunay for his work on this topic from 1934.
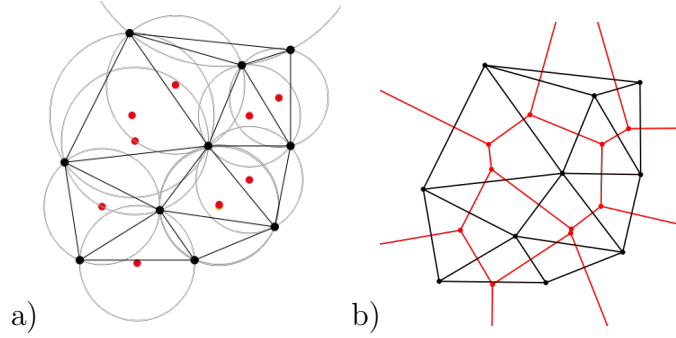
a) The Delaunay triangulation with all the circumcircles and their centers (in red). b) Connecting the centers of the circumcircles produces the Voronoi diagram (in red).

The Delaunay triangulation is used in the DOUAR code which is based on a particle levelset function to track materials. These particles are connected by means of a Delaunay triangulation (usually in a plane at startup, and then in a local Euclidean geometry once the surface is deformed) [144].

 Relevant Literature: [472].

Once a Delaunay triangulation has been obtained it can be used as a FEM mesh. Triangle-based meshes are obviously better suited for simulations of complex geometries:
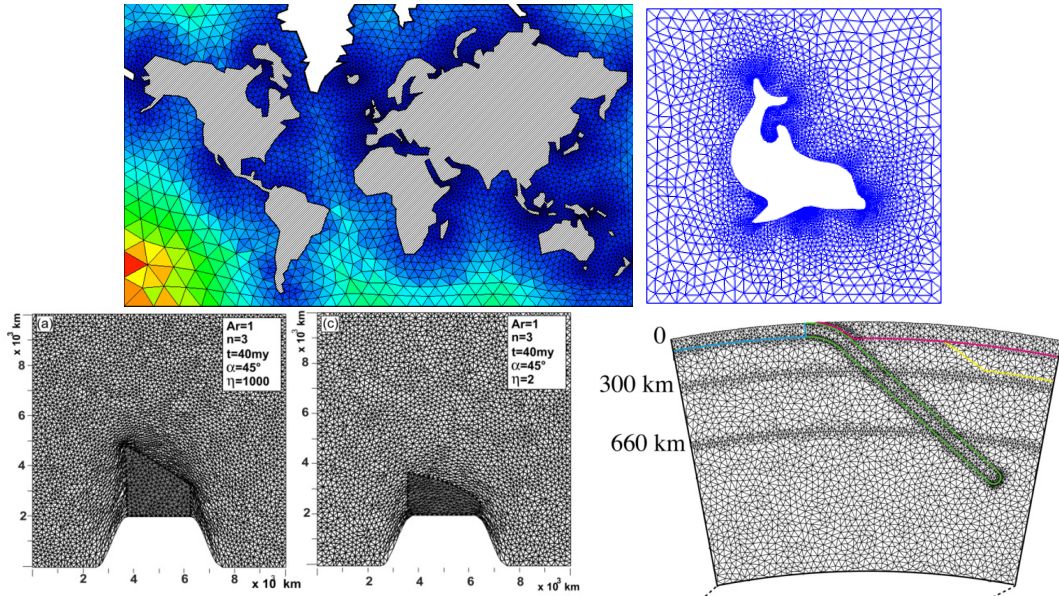
Bottom row. Left: Robl and Stüwe [1153] (2005), Right: Gerault, Becker, Kaus, Faccenna, Moresi, and Husson [474] (2012).

A very practical 2D triangle mesher is the code *Triangle*[12] written by J.R. Shewchuk [1235, 1236, 1237]. Triangle is specialized for creating two-dimensional finite element meshes, but can also perform simpler related tasks such as forming Delaunay triangulations under various assumptions. Another very common mesher tool is Gmsh [485].
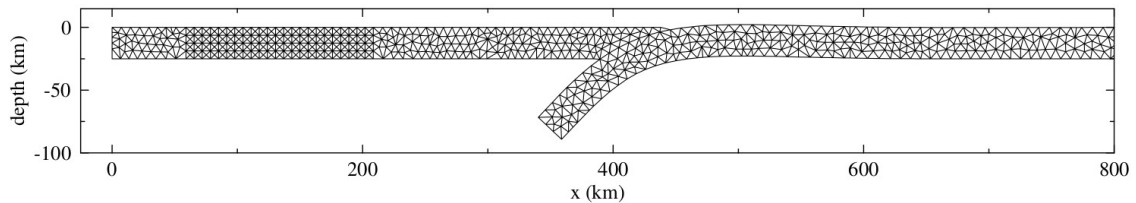
Taken from Buiter *et al.* [176]. Finite element grid. The subducting plate initially extends to 1226 km in the horizontal direction and is not completely shown here. Discretization in the subducting plate is slightly coarser towards the right edge.

---

[12]https://www.cs.cmu.edu/~quake/triangle.html
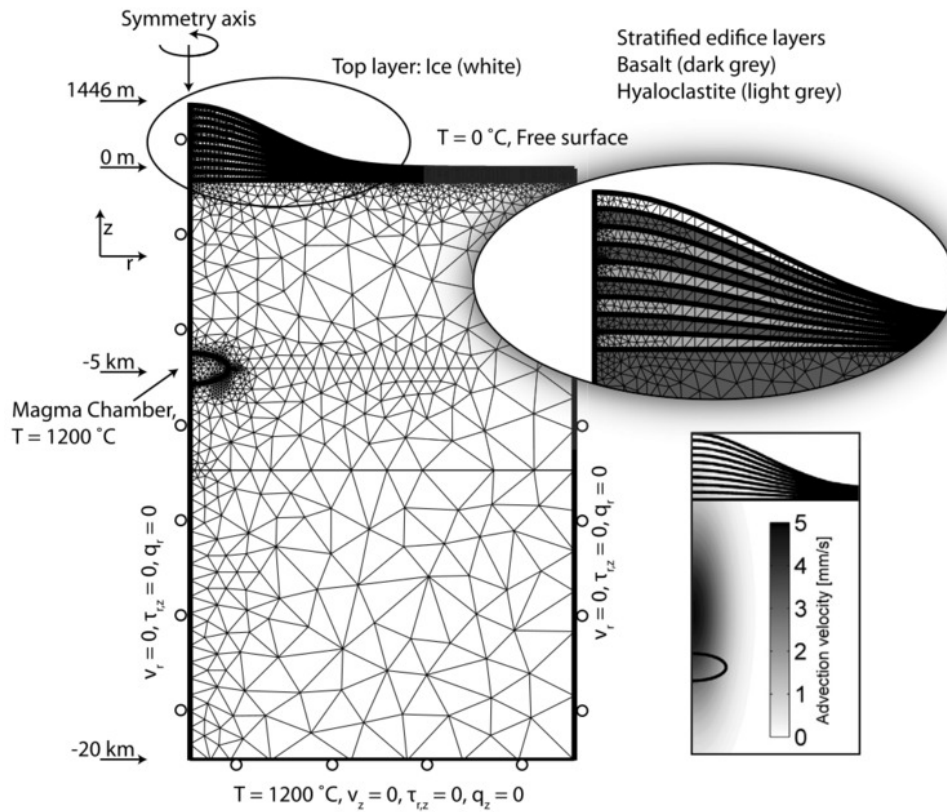
Taken from Bakker, Frehner, and Lupi [43] (2016). Numerical model setup of the 2D axisymmetric half-space with all applied boundary conditions to study the effects of ice-cap unloading on shallow volcanic systems.
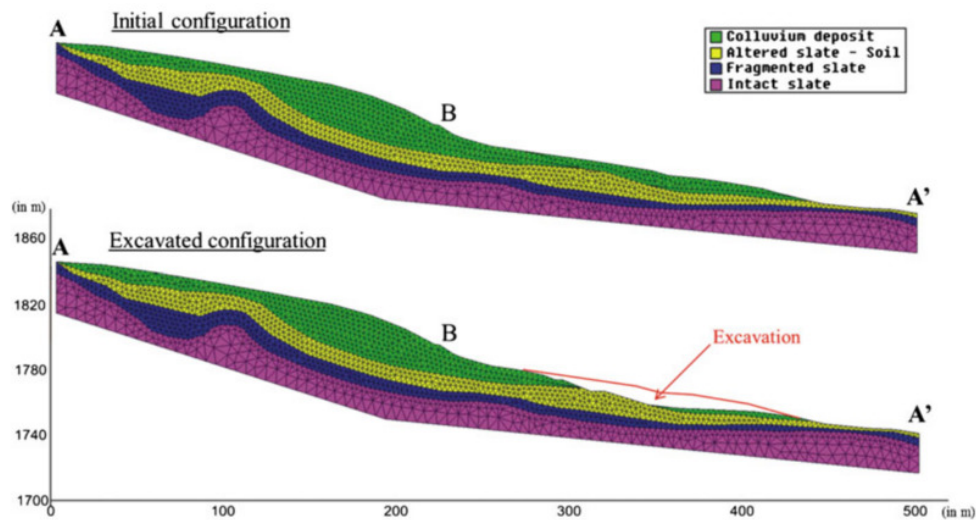
Taken from Fernández-Merodo, García-Davalillo, Herrera, Mira, and Pastor [416] (2014). Modelling of slow landslides. Finite element mesh in the initial and excavated configuration.

Although it is rarely used in practice it is possible to produce meshes which contain both quadrilateral and triangular elements:
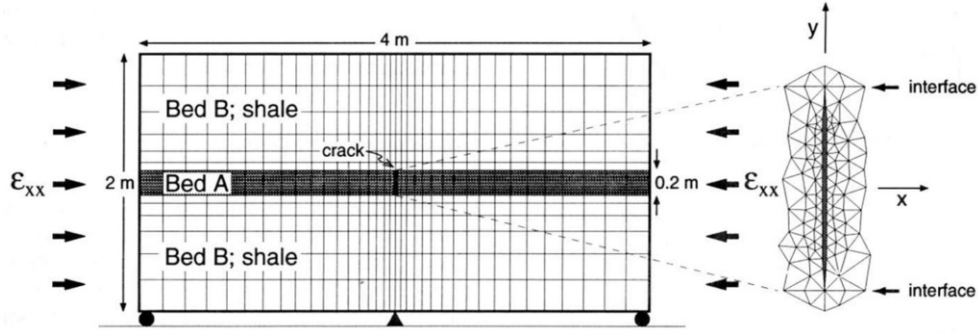
Taken from Fischer, Gross, Engelder, and Greenfield [421]. Mesh used to analayse the stress distribution around a pressurized crack in a layered elastic medium.

Mesh quality In Cioncolini and Boffi [276] (2019) the authors check the mesh quality of their triangulation by computing the following mesaures per element (they also refer to Field [420]):

$$q_1 = \frac{(b+c-a)(c+a-b)(a+b-c)}{abc}$$

$$q_2 = \frac{4\sqrt{3}A_T}{a^2 + b^2 + c^2}$$

where a, b, c are the triangle side lengths and A_T is the triangle area. An equilateral triangle has $q_1 = q_2 = 1$ while a degenerate, zero area triangle has $q_1 = q_2 = 0$. As a rule-of-thumb, in a good quality mesh all triangles should have q_1 , q_2 above about 0.4-0.5.

🗃 Relevant Literature: E. Mulyukova, B. Steinberger, M. Dabrowski, and S.V. Sobolev. "Survival of LLSVPs for billions of years in a vigorously convecting mantle: Replenishment and destruction of chemical anomaly". In: *J. Geophys. Res.: Solid Earth* 120 (2015), pp. 3824–3847. DOI: 10.1002/2014JB011688Mirko Velić, Dave May, and Louis Moresi. "A fast robust algorithm for computing discrete voronoi diagrams". In: *Journal of Mathematical Modelling and Algorithms* 8.3 (2009), pp. 343–355. DOI: 10.1007/s10852-008-9097-6

Remark. *The Natural Neighbour Interpolation method of Sambridge et al. [1174, 1173] is based on the Delaunay triangulation.*

Remark. *Moresi & Mather [957] have released Stripy, a A Python module for (constrained) triangulation in Cartesian coordinates and on a sphere, which is based on Stripack [1127, 1128].*

write about gmesh
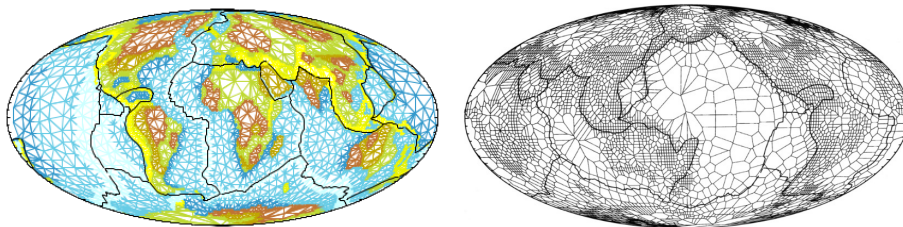
Taken from Gudmundsson & Sambridge (1998) [529]. Boundaries of Voronoi cells around 4100 of the original 16,200 2x2 degree cells selected to sample the details of the regionalization.

16.7.3 Tetrahedra
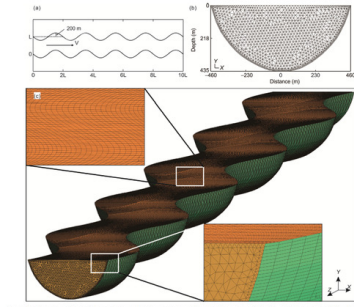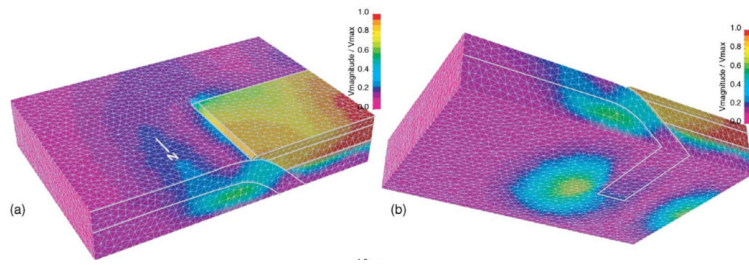
Left: Example of 3D mesh Yang and Shi [1480] (2015); Right: Normalized velocities of a STEP subduction model Govers and Wortel [509] (2005).
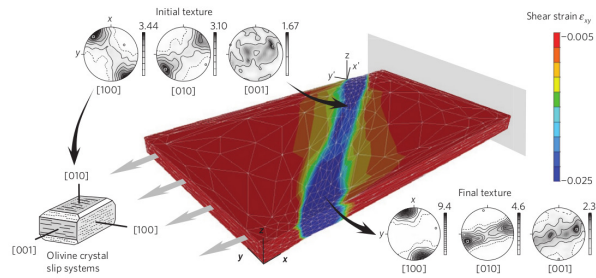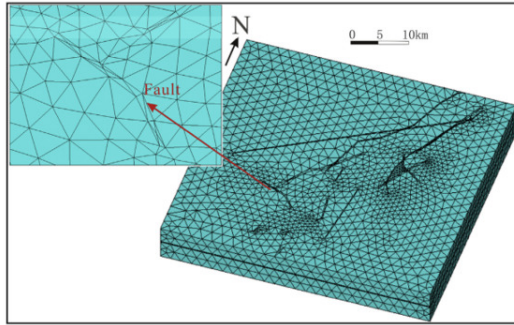
Left: 3D finite element grid in Damintun area, including prescribed faults, Guo, Yao, and Ren [539] (2016); Right: Structural reactivation in plate tectonics controlled by olivine crystal anisotropy, Tommasi, Knoll, Vauchez, Signorelli, Thoraval, and Logé [1360] (2009) - based on $\P_1^+ \times P_1$ elements.
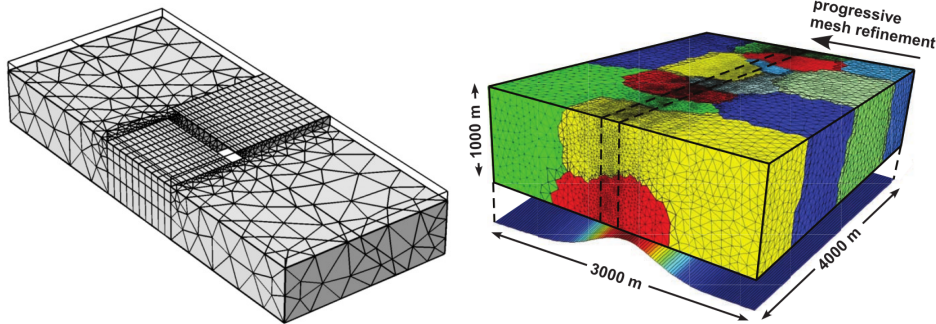
Left: Mesh used for the three-dimensional model. A high resolution mesh is used in the wedge and subslab domains, while the mesh resolution decays to lower values toward the edge of the model. All elements are quadratic, allowing for twice the resolution visualized here, Paczkowski, Montési, Long, and Thissen [1028] (2014); Right: Mid-Ocean Ridge Hydrothermal System: 3D mesh consisting of 2.5 m tetrahedron elements. Resolution is refined toward the axial center, with the finest resolution between the dashed lines, and colors indicate computational domains assigned to separate processors, Coumou, Driesner, and Heinrich [299] (2008).
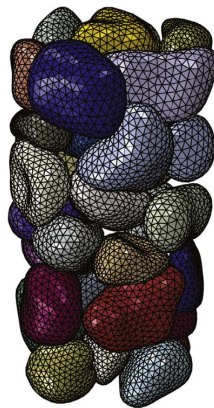
Grains of sand before a compression experiment with FEM Imseeh and Alshibli [654] (2018).

Check TetGen mesher Hang Si. "TetGen, a Delaunay-based quality tetrahedral mesh generator". In: *ACM Transactions on Mathematical Software (TOMS)* 41.2 (2015), pp. 1–36. DOI: 10.1145/2629697.
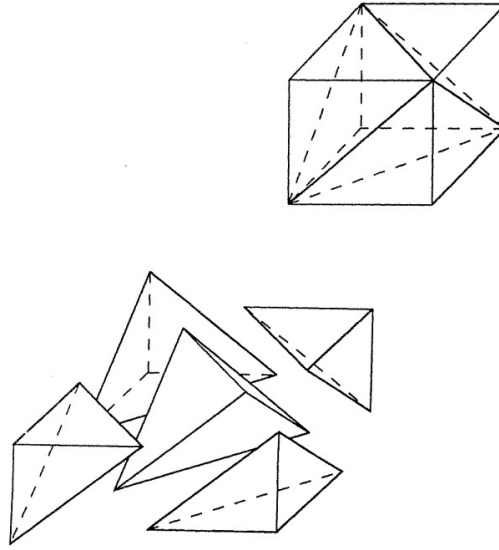
Decomposition of a hexahedron into five tetrahedra. Taken from [89].

16.7.4 Hexahedra

A hexahedron is a convex polytope isomorphic to the cube $[0, 1]^3$. Edges are line segments, facets are strictly **planar** convex polygons.

[Truegrid]

[Ansys]
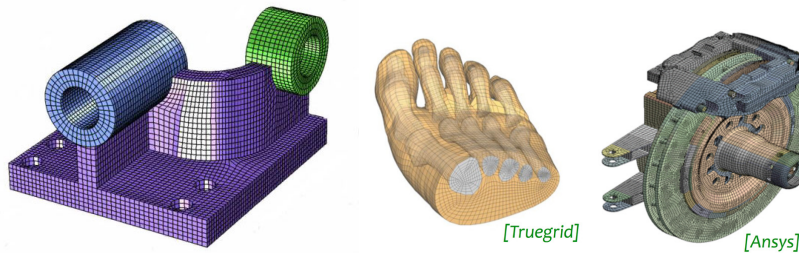
📚 Relevant LiteratureEfficient Volume computation for Three- Dimensional hexahedral Cells [372, 510]

16.7.5 Adaptive Mesh Refinement

Let us do a simple calculation and assume we wish to model mantle convection on Earth. The inner radius is $R_1 = 3485$ km and the bottom of the lithosphere is at $R_2 = 6250$ km. The volume of fluid is then

$$V = \frac{4}{3}\pi(R_2^3 - R_1^3) \simeq 8.5 \times 10^{11} \text{ km}^3$$

Let us further assume that we are satisfied with an average resolution of 10 km. Each element/cell is then 10^3 km^3 and the total number of elements/cell is then

$$N \simeq 8.5 \times 10^8 \sim \mathcal{O}(10^9)$$

This is a very large number. The resulting linear systems from the discretisation of the equations on such a mesh will be very even larger for the Stokes equations and solving these systems will require *very* large numbers of CPUs and long compute times.

Aside from these considerations it is quite obvious that a high resolution mesh is not needed in parts of the mantle where large scale upwellings and downwellings occur, but probably even higher resolution will be needed in the vicinity of thin plumes and boundary layers. This means that a uniform mesh is a sub-optimal way of discretising space for such problems.

The same reasoning also holds in the lithosphere where for instance narrow plate boundaries need to be adequately resolved while the inside of rigid plates can be modelled with coarser meshes.

Finally, although one could employ meshing software to arrive at well balanced meshes in space, the dynamic character of the geodynamics modelling renders this approach cumbersome. A subduction zone, a mid-ocean rift or an ascending plume will evolve in time and the mesh will have to evolve in time too.

In light of all this, it was only a matter of time before Adaptive Mesh Refinement was adopted in computatinal geodynamics. However, since the use and update of such meshes is somewhat complex in terms of numerical algorithms, its introduction came somewhat late (00's and later). The DOUAR code (see Section **??**) developed originally by J. Braun and Ph. Fullsack is a prime example of an early multi-purpose code relying on a self-written Octree library [144]. More recently the ASPECT code was developed on top of the Octree library p4est [205]. Note the 2007 and 2008 papers by Davies et al [326, 329] which explore adaptive mesh refinement with the ConMan code (see Appendix **??**).

For further reading I suggest you read the review by May, Schellart & Moresi on this topic [896].
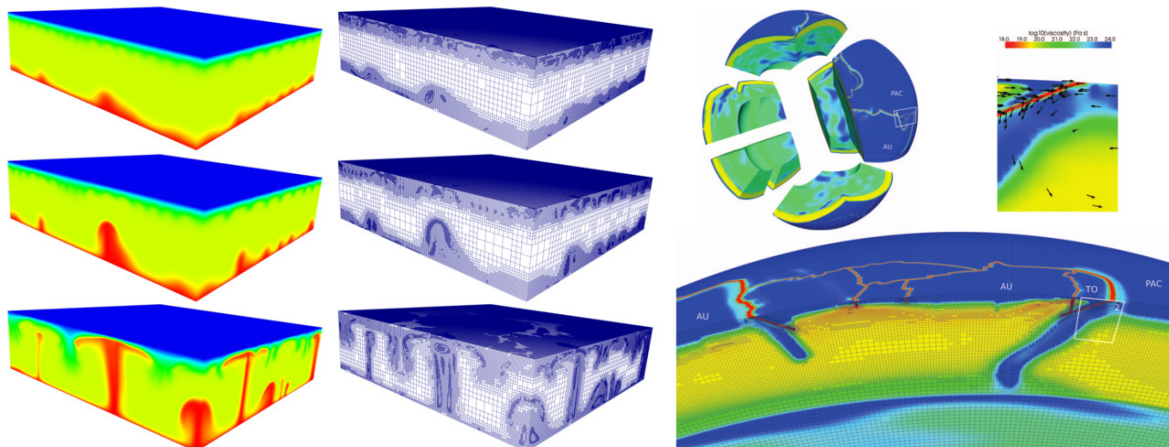
Taken from [204] and [206]
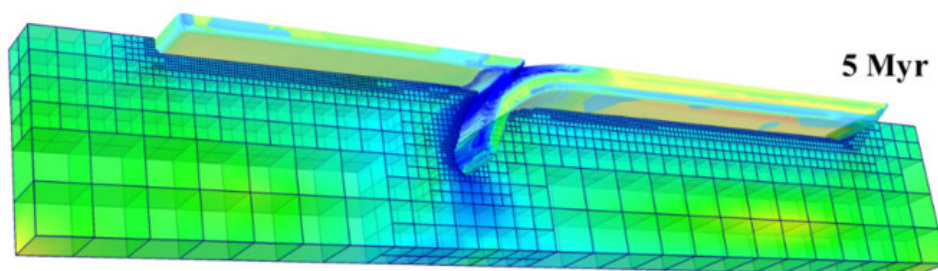
Taken from [496]

📥 Relevant Literature:

- C. Burstedde et al. "Scalable Adaptive Mantle Convection Simulation on Petascale Supercomputers". In: *ACM/IEEE SC Conference Series, 2008* (2008)

- C. Burstedde, O. Ghattas, G. Stadler, T. Tu, and L.C. Wilcox. "Parallel scalable adjoint-based adaptive solution of variable-viscosity Stokes flow problems". In: *Computer Methods in Applied Mechanics and Engineering* 198 (2009), pp. 1691–1700. DOI: 10.1016/j.cma.2008.12.015

- Carsten Burstedde et al. "Extreme-scale AMR". in: *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society. 2010, pp. 1–12. DOI: `10.1109/SC.2010.25`

- W. Leng and S. Zhong. "Implementation and application of adaptive mesh refinement for thermochemical mantle convection studies". In: *Geochem. Geophys. Geosyst.* 12.4 (2011). DOI: `10.1029/2010GC003425`

- Y. Mishin. "Adaptive multiresolution methods for problems of computational geodynamics". PhD thesis. ETH Zurich, 2011

- K. Sverdrup, N. Nikiforakis, and A. Almgren. "Highly parallelisable simulations of time-dependent viscoplastic fluid flow simulations with structured adaptive mesh refinement". In: *Physics of Fluids* 30 (2018), p. 093102. DOI: `10.1063/1.5049202`

- Marc Fehling and Wolfgang Bangerth. "Algorithms for parallel generic hp-adaptive finite element software". In: *ACM Transactions on Mathematical Software* 49.3 (2023), pp. 1–26. DOI: `10.1145/3603372`

A short illustrative exercise .
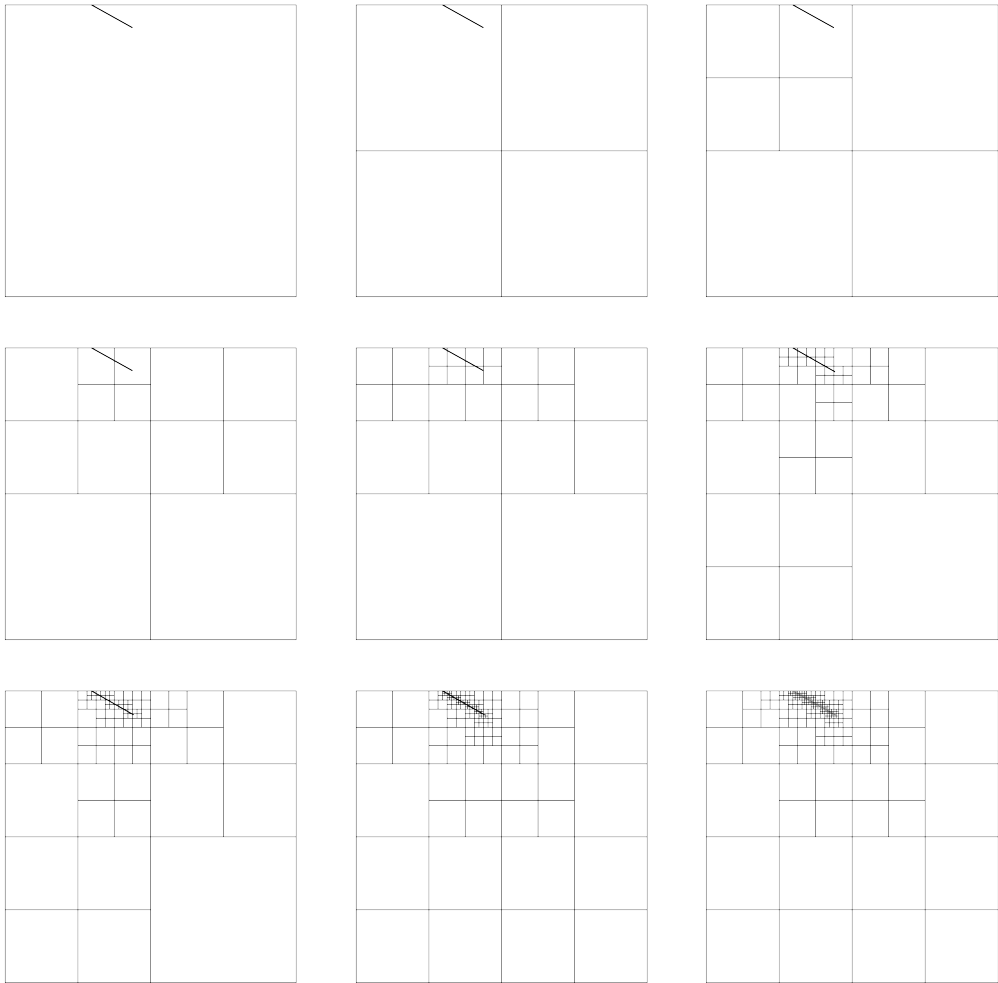
	# l0	# l1	# l2	# l3	# l4	# l5	# l6	# l7	# l8
max level= 0	1								
max level= 1	0	4							
max level= 2	0	3	4						
max level= 3	0	2	7	4					
max level= 4	0	2	5	10	8				
max level= 5	0	1	8	12	11	20			
max level= 6	0	1	8	11	13	20	32		
max level= 7	0	0	11	14	15	23	37	60	
max level= 8	0	0	11	13	17	27	43	72	116
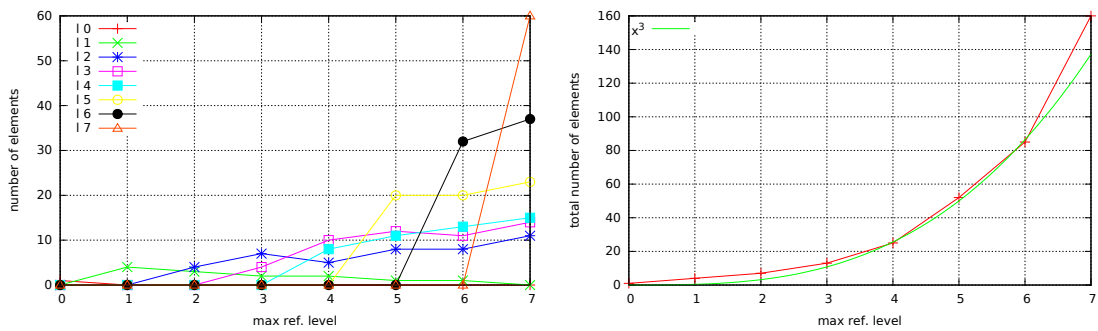
In the particular case presented here, even though the inclusion in a short two-dimensional line, the total number of elements grows faster than the third power of the refinement level. While of course the total number of elements remains much smaller than the constant resolution counterpart,

this observation tells us that authorising a unit increase of the maximum refinement level can have a substantial effect on the total number of elements.
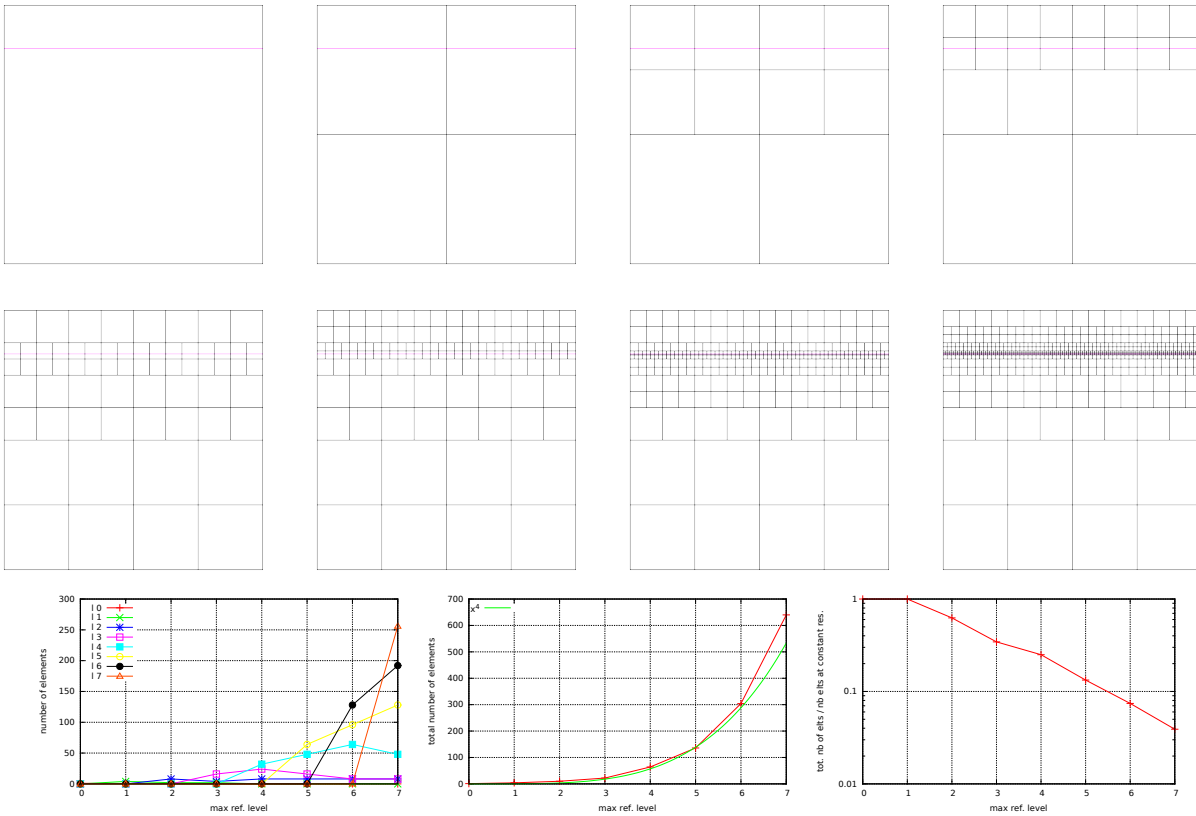
16.7.6   Conformal Mesh Refinement

The quadtree/octree mesh refinement presented above is one option when it comes to mesh refinement (or h-refinement). However their massive drawback is the presence of hanging notes which require special attention. Another approach to mesh refinement is conformal mesh refinement as best exemplified on the following figures:
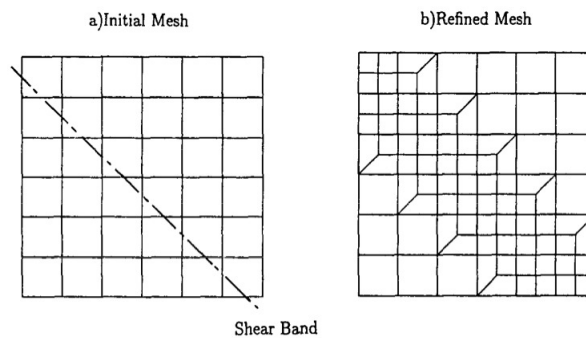
Taken from Deb *et al.* (1996) [343]. A typical instance of the outcome of the refinement procedure. Notice that the 'spill-over' is reduced to one row on each side of the 'localized' elements.
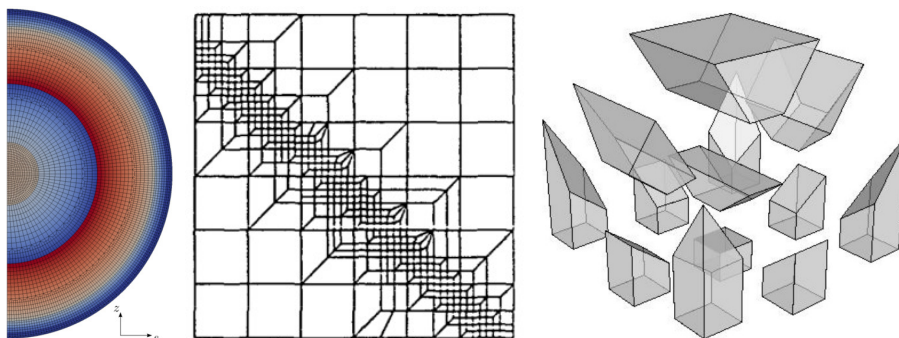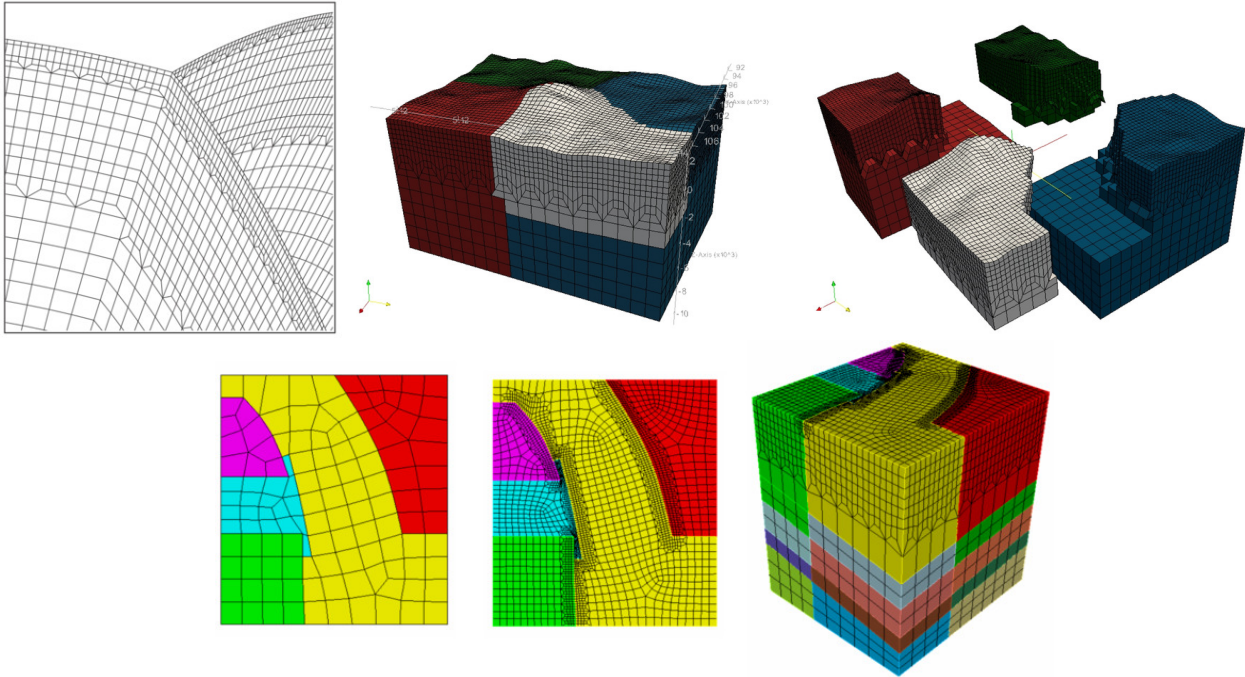
Top row, From left to right: van Driel *et al.* (2015) [368]; Deb *et al.* (1996) [343]; Harris *et al.* (2004) [583]; Komatitsch *et al.* (2005) [763]; Middle row: Specfem manual; Bottom row: I don't know anymore.
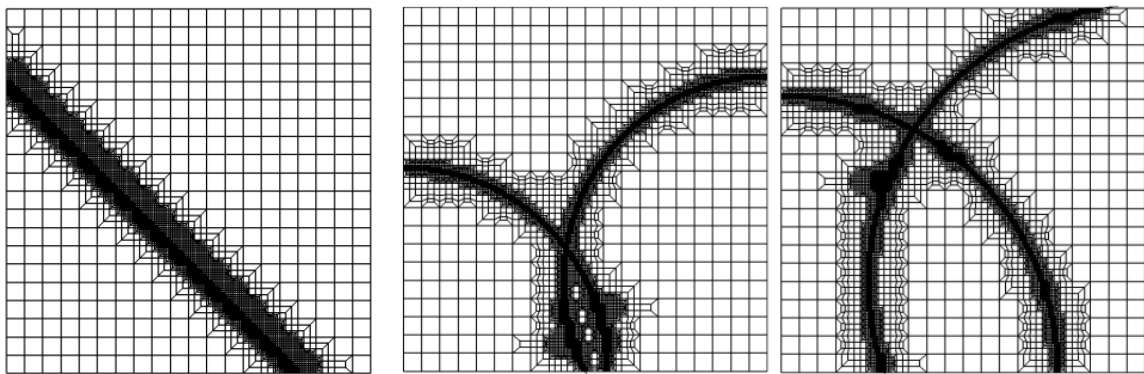
Taken from Garimella [462] (2009).

[Ito Shih Soni '08]   [Maréchal '09]
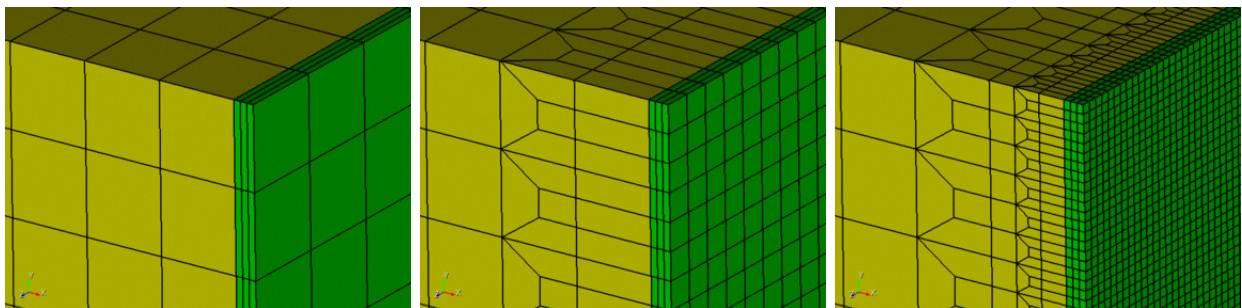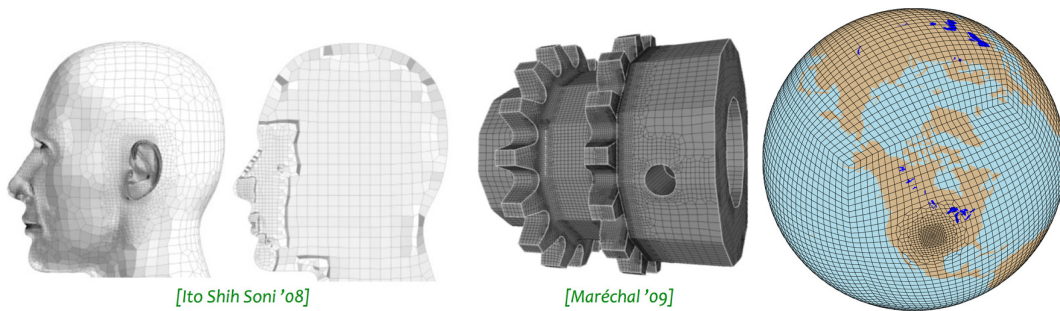
https://cubit.sandia.gov/public/14.0/help_manual/WebHelp/mesh_generation/mesh_
modification/mesh_refinement.htm

📚 Relevant Literature:

- Düster & Rank [377],

- Harris *et al.* (2004) [583],

- Anderson *et al.* (2009) [20],

- Anderson [19],

- Garimella (2009) [462],

- Nicolas & Fouquet (2013) [997, 996].

- Parrish [1039],

- Schneiders [1208, 1207, 1210, 1209],

- Schneiders *et al.* [1211],

- Staten & Canann [1275],

- book by Ramm *et al.* [1106].

16.7.7 Stretching the mesh

In some cases the topology of the mesh can be regular but one can for instance stretch the mesh such that (for instance) the vertical resolution is higher at the top than at the bottom, or higher in the middle than on the sides.

The idea behind the transformation is a piecewise-linear function which maps [0,L] to [0,L] where L is the length of the domain in the x-direction. For instance, this transformation can take the following form:
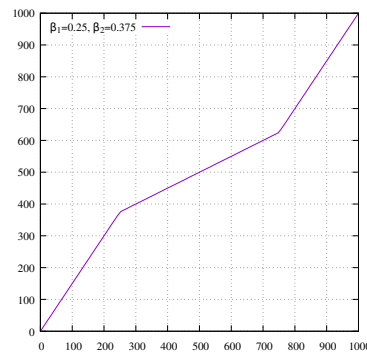
Parameters β_1 and β_2 control the shape of the lines.

The kinks in the line occur at $\beta_1 L$ and $(1 - \beta_1)L$ (see code here under).

The (minimal) code to transform the mesh is as follows:

```
def stretch_towards_center(x,L,beta1,beta2):
    if x<beta1*L:
        val = beta2/beta1*x
    elif x<(1.-beta1)*L:
        val = (1-2*beta2)/(1-2*beta1)*(x-beta1*L)+beta2*L
    else:
        val=beta2/beta1*(x-(1-beta1)*L)+(1-beta2)*L
    return val
```

```
[...]

beta1=0.25
beta2=0.375

for i in range(0,NV):
    x[i]=stretch_towards_center(x[i],Lx,beta1,beta2)
```

The following meshes count 64x16 elements. The top one is a regular mesh, with square elements, while the second one has been stretched by means of the transformation above:
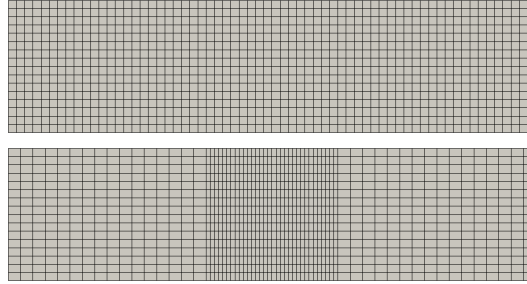
Concerning the stretching towards the top of the model domain, the transformation line is as follows:
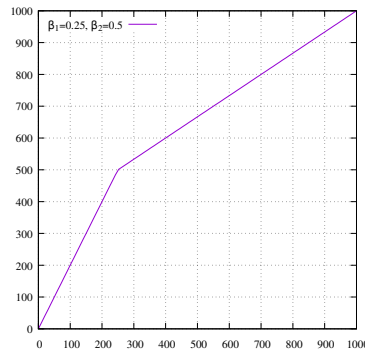
Parameters β_1 and β_2 control the shape of the lines. The kinks in the line occur at $\beta_1 L$ and $(1 - \beta_1)L$.

The slope of the left line is $\beta_2/\beta_1 x$.

The (minimal) code to transform the mesh is as follows:

```
def stretch_towards_top(x,L,beta1,beta2):
    if x<beta1*L:
        val=beta2/beta1*x
    else:
        val=(1-beta2)/(1-beta1)*(x-beta1*L)+beta2*L
    return val

[...]

beta1=0.25
beta2=0.5
for i in range(0,NV):
    y[i]=stretch_towards_top(y[i],Ly,beta1,beta2)
```

The following meshes count 64x16 elements. The top one is a regular mesh, with square elements, while the second one has been stretched by means of the transformation above.

729
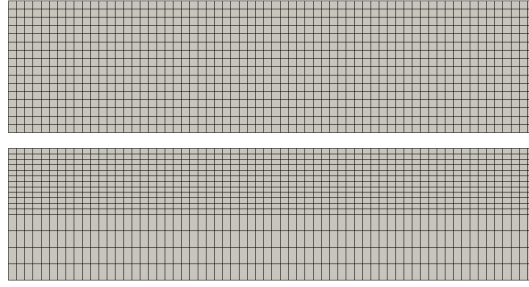
Finally both transformations can be applied to the same mesh:
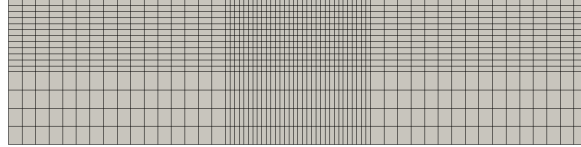
This approach is used in STONE 67.

16.7.8   Meshes in an annulus
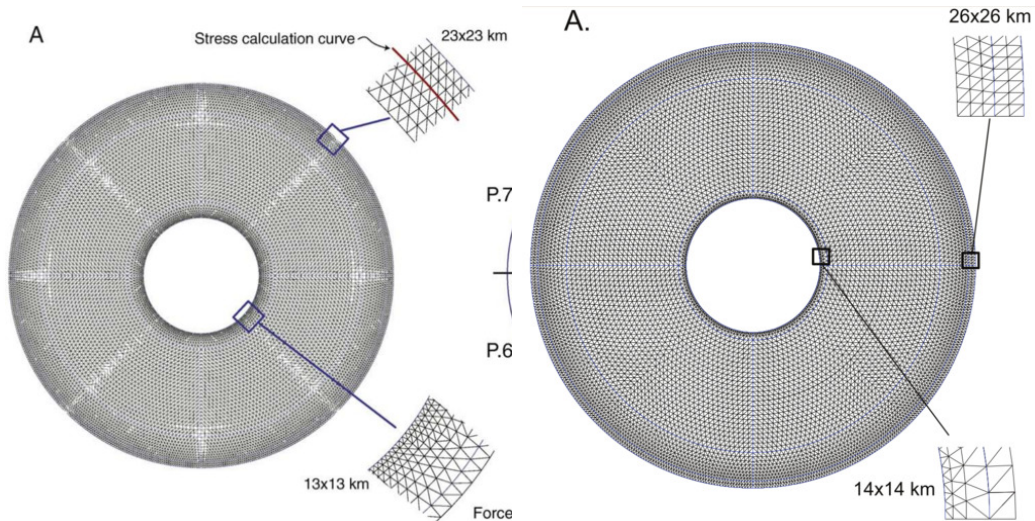
The quadratic finite element mesh as used in Brandenburg *et al.* [141, 142]

16.7.9   Meshes in/on a hollow sphere

The following is for the most part published in Thieulot (2018) Thieulot [1344] (2018).

To a first approximation the Earth is a sphere: the Earth's polar diameter is about 43 kilometers shorter than its equatorial diameter, a negligible difference of about 0.3%. As a consequence, modelling physical processes which take place in the planet require the discretisation of a sphere. Furthermore, because core dynamics occur on vastly difference time scales than mantle dynamics, mantle modelling usually leaves the core out, thereby requiring simulations to be run on a hollow sphere mesh (with the noticeable exception of Gerya and Yuen [478]).

Although so-called latitude-longitude grids would seem appealing, they suffer from the convergence of meridians at the poles (resulting in over sampling at poles) and the juxtaposition of triangles near the poles and quadrilaterals elsewhere. As a consequence more regular, but more complex, grids have been designed over the years which tesselate the surface of the sphere into triangles or quadrilaterals (sometimes overlapping). There is the 'cubed sphere' [1156, 597, 253, 1287, 254, 156, 1486], the Yin-Yang grid [701, 1488, 1490, 703, 1312, 302, 303], the Yin-Yang-zhong grid [588], the Yin-yang grid of Shahnas & Peltier [1229], the spiral grid [649], an icosahedron-based grid [63, 872], or

a grid composed of 12 blocks further subdivided into quadrilaterals [1523] as used in the CitcomS code. Note that [1013] have also presented a method for generating a numerical grid on a spherical surface which allows the grid to be based on several different regular polyhedrons (including octahedron, cube, icosahedron, and rhombic dodecahedron). Ideally, one wishes to generate a mesh that is regular, i.e. angles between edges/faces as close to 90° as possible, of approximately similar volumes.
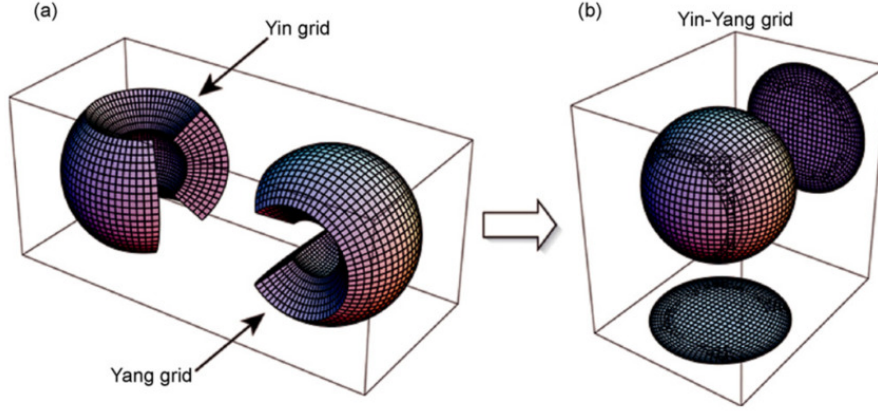
Example of Yin-Yang grid. Taken from Kameyama, Kageyamab, and Sato [703] (2008).

How such meshes are built is often not discussed in the literature. It is a tedious exercise of three-dimensional geometry and it can be time-consuming, especially the connectivity array generation. In Thieulot (2018) [1344] I present an open source mesh generator for three hollow sphere meshes: the 'cubed sphere' mesh, the CitcomS mesh and the icosahedral mesh:

- The cubed sphere ('HS06'), composed of 6 blocks which are themselves subdivided into $N_b \times N_b$ quadrilateral shaped cells [1169, 1156, 597, 203]. Four types of cubed spheres meshes have been proposed: the conformal, elliptic, gnomonic and spring types [1087]:
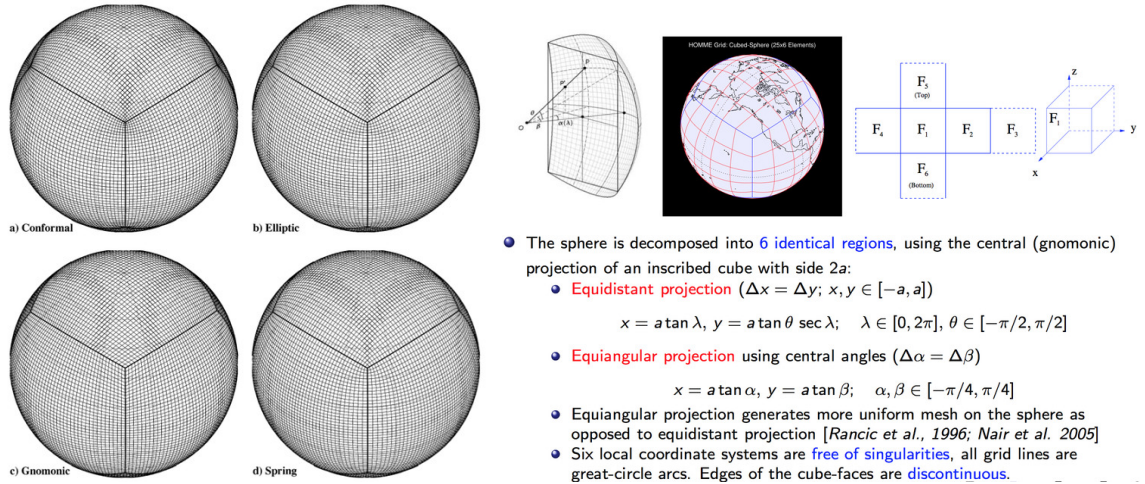
Left: The cubed-sphere grids at 2° resolution displaying cells on the sphere, the image focuses on the distribution of grid cells near one corner of the grid; (a) conformal mapping [1111, 899], (b) the gnomonic grid modified by elliptic solver, (c) equiangular gnomonic mapping and (d) the gnomonic grid modified by spring dynamics. [1087]. Right: Taken from presentation by R. Nair, see Nair2008.pdf

However only gnomonic meshes are considered in Thieulot (2018): these are obtained by inscribing a cube within a sphere and expanding to the surface of the sphere. The cubed sphere has been used in large-scale mantle convection simulation in conjunction with Adaptive Mesh Refinement [7, 203].

731
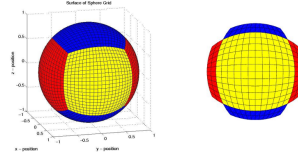
- The CitcomS mesh ('HS12') composed of 12 blocks also subdivided into $N_b \times N_b$ quadrilateral shaped cells [1523, 1287, 1521, 30]. Note that ASPECT [776, 593], a relatively new code aimed at superseeding CitcomS can generate and use this type of mesh [1341] but is not limited to it.
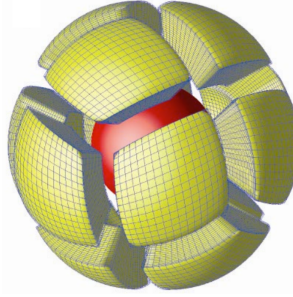
- The icosahedral mesh ('HS20') composed of 20 triangular blocks [63, 62] subdivided into triangles, which is used in the TERRA code [188, 187, 186, 327].
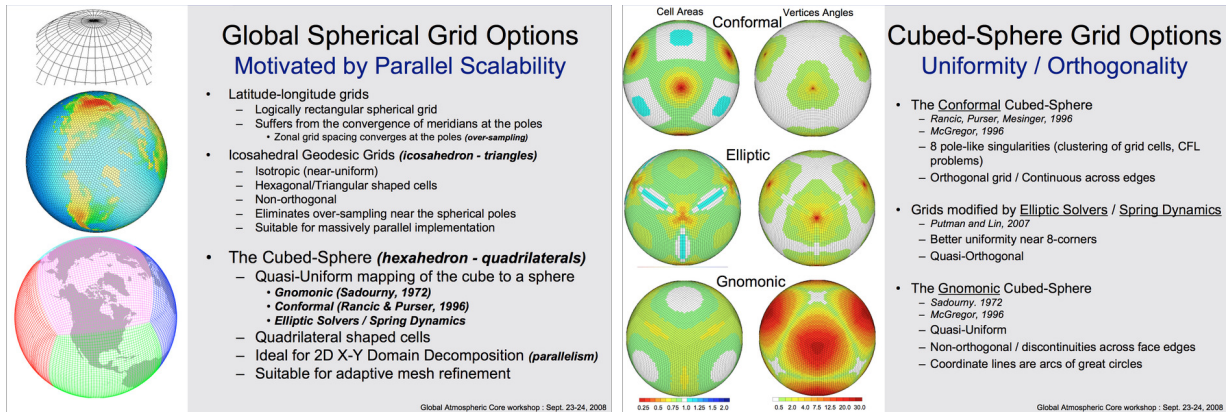
source?

Given the regularity and symmetry of these meshes determining the location of the mesh nodes in space is a relatively straightforward task. Building the mesh connectivity in an efficient manner is where the difficulty lies.

The approach to building all three meshes is identical:

1. A reference square or triangle is populated with cells, parametrised by a level l: the square is subdivided into $l \times l$ quadrilaterals while the triangle is subdivided into l^2 triangles.
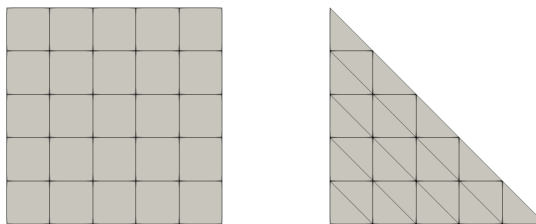
Reference square and triangles meshes at level 5.

2. This reference square or triangle is then replicated *nblock* times (6, 12 or 20) and mapped onto a portion of a unit sphere. The blocks are such that their union covers a full sphere but they cannot overlap except at the edges:
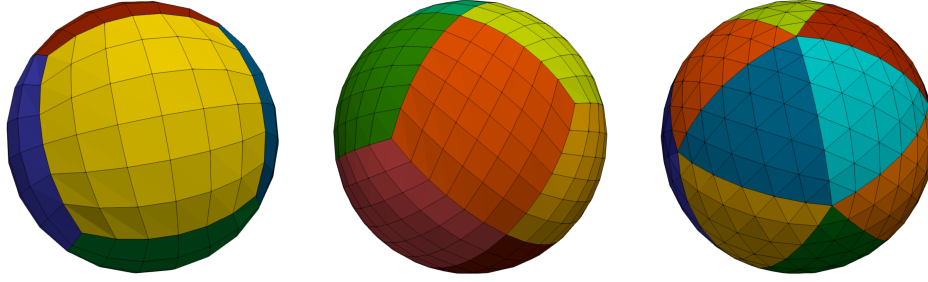
From left to right: HS06, HS12 and HS20 shells coloured by block number.

3. All block meshes are then merged together to generate a shell mesh. This task is rather complex as duplicate nodes must be removed and all connectivity arrays of the blocks must then be mended accordingly.

4. Shell meshes are replicated *nlayer+1* times outwards with increasing radii. The *nlayer* shells are then merged together to form a hollow sphere mesh:
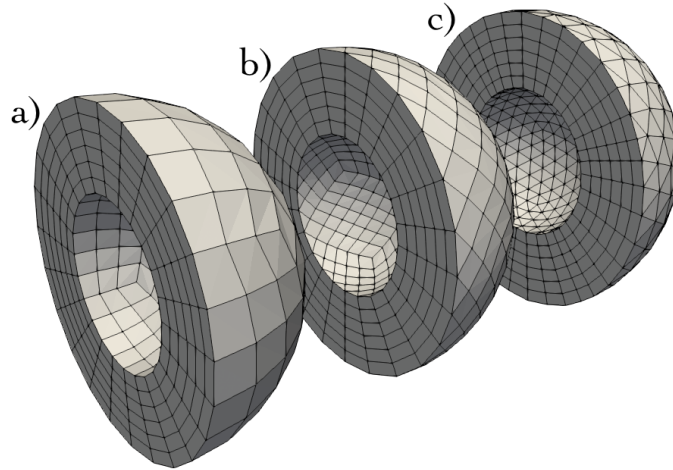
a) HS06 mesh composed of 6 blocks containing each 6^3 cells; b) HS12 mesh composed of 12 blocks containing each 6^3 cells; e) HS20 mesh composed of 20 blocks containing each 6^3 cells.

More information on these steps is available in the manual of the code. In the following table the number of nodes and cells for a variety of resolutions for all three mesh types is reported. Looking at the CitcomS literature of the past 20 years, we find that the mesh data presented in this table cover the various resolutions used, e.g. 12×48^3 [908, 30], 12×64^3 [179] 12×96^3 [180], 12×128^3 [70, 1448, 1449]. Note that in the case of the HS06 and HS12 meshes the mesh nodes are mapped out to the 6 or 12 blocks following either an equidistant or equiangle approach (see [1087] for details on both approaches).

type	level	N	N_{el}	structure
HS06	2	78	48	6×2^3
HS06	4	490	384	6×4^3
HS06	8	3,474	3,072	6×8^3
HS06	16	26,146	24,576	6×16^3
HS06	32	202,818	196,608	6×32^3
HS06	64	1,597,570	1,572,864	6×64^3
HS06	128	12,681,474	12,582,912	6×128^3
HS06	256	101,057,026	100,663,296	6×256^3
HS12	2	150	96	12×2^3
HS12	4	970	768	12×4^3
HS12	8	6,930	6,144	12×8^3
HS12	16	52,258	49,152	12×16^3
HS12	32	405,570	393,216	12×32^3
HS12	48	1,354,850	1,327,104	12×48^3
HS12	64	3,195,010	3,145,728	12×64^3
HS12	128	25,362,690	25,165,824	12×128^3
HS12	256	202,113,538	201,326,592	12×256^3
HS20	2	126	160	20×2^3
HS20	4	810	1,280	20×4^3
HS20	8	5,778	10,240	20×8^3
HS20	16	43,554	81,920	20×16^3
HS20	32	337,986	655,360	20×32^3
HS20	64	2,662,530	5,242,880	20×64^3
HS20	128	21,135,618	41,943,040	20×128^3
HS20	256	168,428,034	335,544,320	20×256^3

Number of nodes N and elements/cells N_{el} for the three types of meshes and for various levels.

HS06: cubed sphere; HS12: CitcomS mesh; HS20: icosahedral mesh.

There are also many possibilities offered by the use of tetrahedral cells/elements:

Left: Grid of a global neo-tectonic SHELLS model coupled to a global mantle circulation model; colours represent temperatures (red=hot, blue=cold) at a depth of 200km below the surface. Taken from Oeser *et al.* (2009) [1009]. Right: Taken from the GeoTess software [13] manual.

[13] https://www.sandia.gov/geotess/
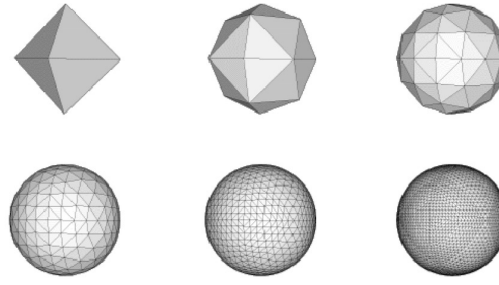
Example of a Hierarchical Triangular Mesh
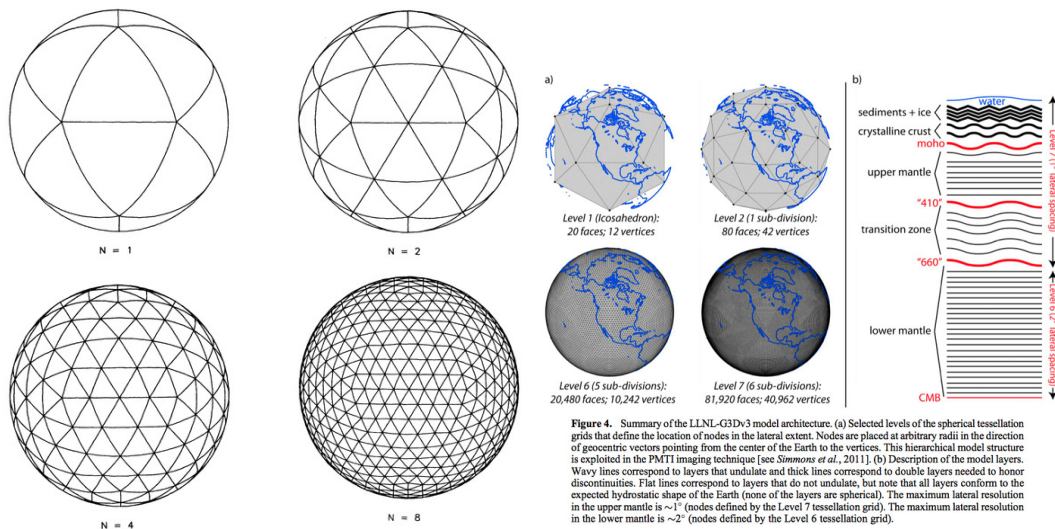
Baumgardner and Frederickson [63] (1985), Simmons, Myers, Johannesson, and Matzel [1249] (2012)

📚 Relevant Literature:

- Phillips, Davies, and Oldham [1058] (2019) present an algorithm which builds polyhedral-based grids.

- Upadhyaya, Sharma, Mittal, and Fatima [1389] (2011) on icosahedral-hexagonal grids on a sphere for CFD applications.

- Saff and Kuijlaars [1170] (1997) on distributing many points on a sphere.

- Swinbank and Purser [1306] (2006) on Fibonacci grids which possess virtually uniform and isotropic resolution, with an equal area for each grid point.

- Hardin, Saff, et al. [581] (2004) on discretizing manifolds via Minimum Energy Points. Element Software

16.8 Pressure smoothing/filtering/recovery for $Q_1 \times P_0$ elements

It has been widely documented that the use of the $Q_1 \times P_0$ element is not without problems. Aside from the consequences it has on the FE matrix properties, we will here focus on another unavoidable side effect: the spurious pressure checkerboard modes.

These modes have been thoroughly analysed decades ago, see for instance Hughes, Liu, and Brooks [641] (1979), Sani, Gresho, Lee, and Griffiths [1184] and Sani, Gresho, Lee, Griffiths, and Engelman [1185] (1981), Griffiths and Silvester [524] (1994). They can be filtered out (Chen, Pan, and Chang [238] (1995)) or simply smoothed (Lee, Gresho, and Sani [804] (1979)), as we will see later. Nodes on edges and corners may need special treatment as documented in Sani, Gresho, Lee, and Griffiths [1184] (1981) or Lee, Gresho, and Sani [804] (1979). The list of 8 schemes is not exhaustive with regards to the above mentioned publications. There has been considerable amount of work on the topic and this section is unfortunately not representing the literature appropriately.

On the following figure (a,b), pressure fields for the lid driven cavity experiment are presented for both an even and un-even number of elements. We see that the amplitude of the modes can sometimes be so large that the 'real' pressure signal is not visible under the checkerboard and that something as simple as the number of elements in the domain can trigger those or not at all.

a)
b)

c)
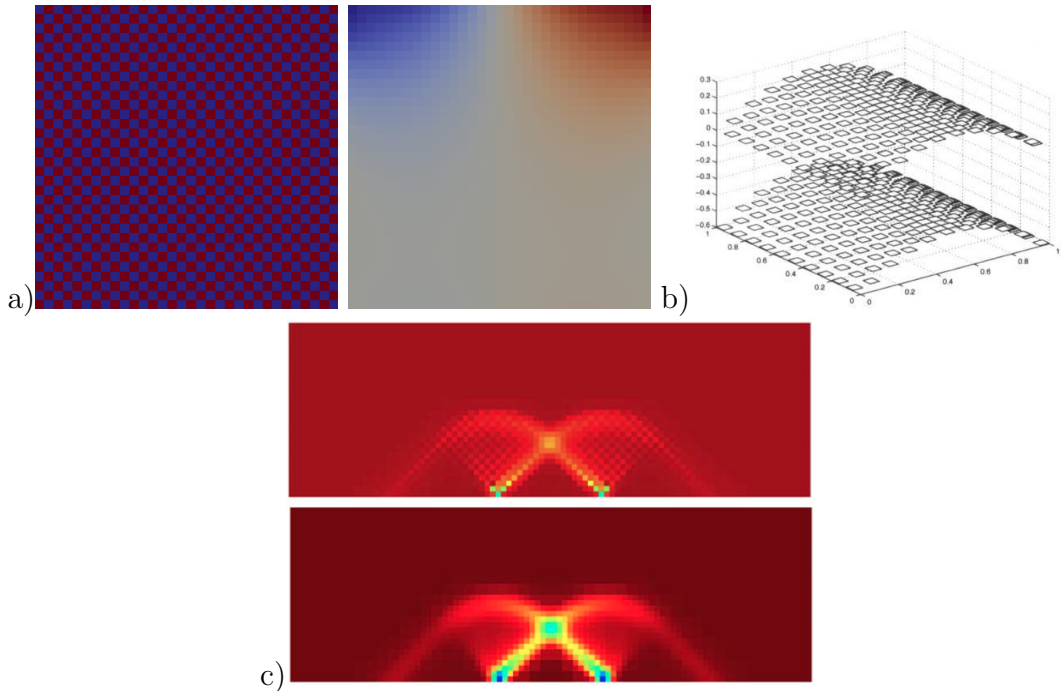
a) element pressure for a 32x32 grid and for a 33x33 grid;

b) image from [363, p307] for a manufactured solution; c) elemental pressure and smoothed pressure for the punch experiment [1347]

16.8.1 Scheme 1

The easiest post-processing step that can be used (especially when a regular grid is used) is explained in Thieulot, Fullsack, and Braun [1347] (2008): "The element-to-node interpolation is performed by averaging the elemental values from elements common to each node; the node-to-element interpolation is performed by averaging the nodal values element-by-element. This method is not only very efficient but produces a smoothing of the pressure that is adapted to the local density of the octree. Note that these two steps can be repeated until a satisfying level of smoothness (and diffusion) of the pressure field is attained."
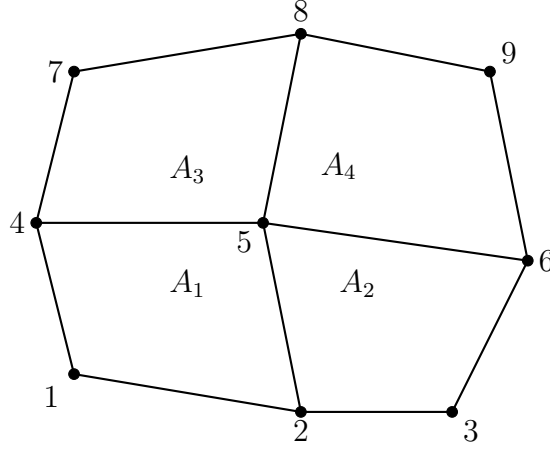
$$q_5^{(1)} = \frac{1}{4} \sum_{e=1}^{4} p_e$$

In the codes which rely on the $Q_1 \times P_0$ element, the (elemental) pressure is simply defined as

```
p=np.zeros(nel,dtype=np.float64)
```

while the nodal pressure is then defined as[14]

```
q=np.zeros(nnp,dtype=np.float64)
```

The element-to-node algorithm is then simply (in 2D):

```
count=np.zeros(nnp,dtype=np.int32)
for iel in range(0,nel):
    q[icon[0,iel]]+=p[iel]
    q[icon[1,iel]]+=p[iel]
    q[icon[2,iel]]+=p[iel]
    q[icon[3,iel]]+=p[iel]
    count[icon[0,iel]]+=1
    count[icon[1,iel]]+=1
    count[icon[2,iel]]+=1
    count[icon[3,iel]]+=1
q=q/count
```

16.8.2   Schemes 2,3

Schemes 2,3 are very similar and are presented in Sani, Gresho, Lee, and Griffiths [1184] and Sani, Gresho, Lee, Griffiths, and Engelman [1185] (1981). Scheme 2 uses the areas of the surrounding elements as weights for the arithmetic averaging while scheme 3 uses the area of the triangles:

---
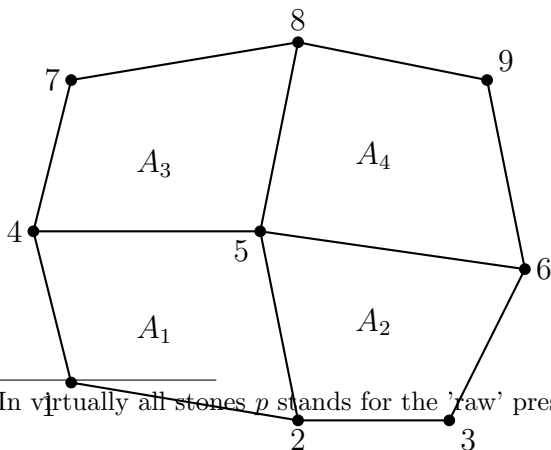
[14]In virtually all stones p stands for the 'raw' pressure and q stands for its projection onto the velocity mesh.
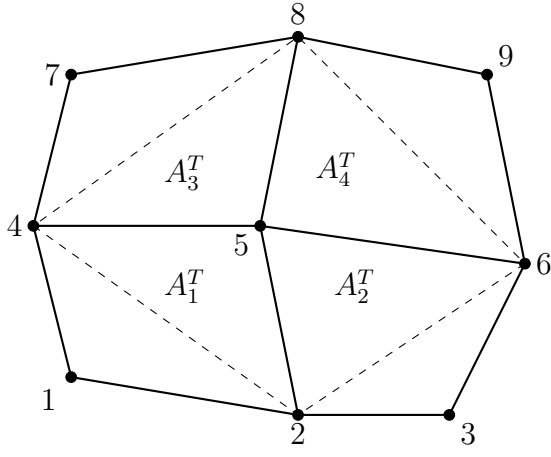
$$q_5^{(2)} = \frac{\sum\limits_{e=1}^{4} A_e p_e}{\sum\limits_{e=1}^{4} A_e} \qquad\qquad q_5^{(3)} = \frac{\sum\limits_{e=1}^{4} A_e^T p_e}{\sum\limits_{e=1}^{4} A_e^T}$$

Remark. *Although Schemes 1,2,3 are similar, scheme 1 is the simplest and fastest to implement since the areas of neighbouring elements/triangles are not needed.*

Remark. *Schemes 1,2,3 are identical if all elements are rectangles of identical dimensions.*

16.8.3   Scheme 4

This scheme has been designed by me. It resembles the last three ones, but the weighing is in this case different.

Let us consider a 1D problem:
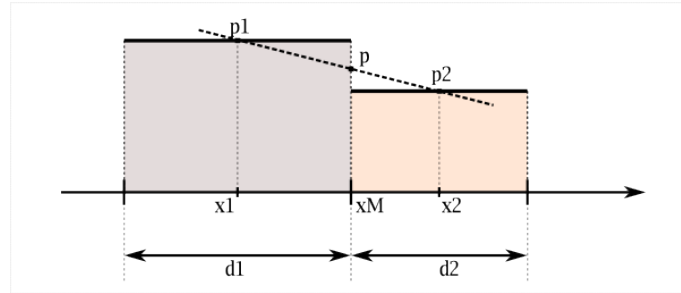
Elemental pressures p_1 and p_2 corresponding to elements 1 and 2 respectively are known at locations x_1 and x_2. The two elements have a different size, characterised in this case by the distances d_1 and d_2 to their common edge.

The equation of the line passing through points (x_1, p_1) and (x_2, p_2) is

$$p(x) = \frac{p_2 - p_1}{x_2 - x_1}(x - x_1) + p_1$$

The x coordinate of the common edge is given by $x = x_1 + d_1/2$, and since $x_2 - x_1 = (d_1 + d_2)/2$, the pressure at this location writes:

$$p(x_M) = \frac{p_2 - p_1}{d_1 + d_2}d_1 + p_1 = \frac{\frac{p_1}{d_1} + \frac{p_2}{d_2}}{\frac{1}{d_1} + \frac{1}{d_2}}$$

738

Extrapolating this formula to 2D, d_1 and d_2 are in fact the element volumes, so that

$$q_5^{(4)} = \frac{\sum\limits_{j=1}^{4} \frac{p_j^e}{A_j^e}}{\sum\limits_{j=1}^{4} \frac{1}{A_j^e}} = \frac{\frac{p_1^e}{A_1^e} + \frac{p_2^e}{A_2^e} + \frac{p_3^e}{A_3^e} + \frac{p_4^e}{A_4^e}}{\frac{1}{A_1^e} + \frac{1}{A_2^e} + \frac{1}{A_3^e} + \frac{1}{A_4^e}}$$

There remains a problem, due to the presence of the boundary nodes for which the sums present in the above equation do not run up to 4. A boundary node only has three neighbours and a corner node only two. Additional measures are required for these nodes.
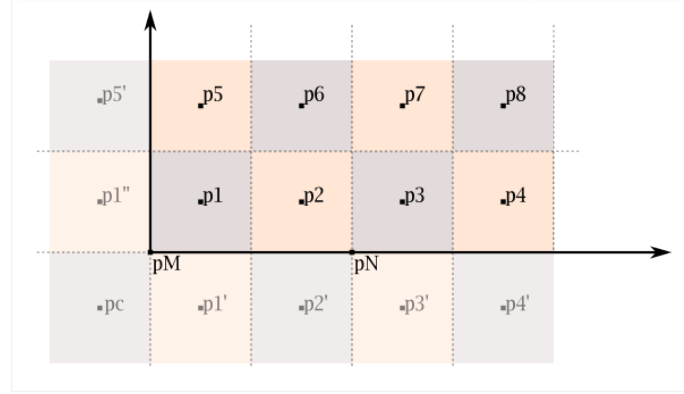
The pressure value p_N is obtained as follows:

$$q_N = \frac{\frac{p_2^e}{A_2^e} + \frac{p_3^e}{A_3^e} + \frac{p_{2'}^e}{A_{2'}^e} + \frac{p_{3'}^e}{A_{3'}^e}}{\frac{1}{A_2^e} + \frac{1}{A_3^e} + \frac{1}{A_{2'}^e} + \frac{1}{A_{3'}^e}}$$

The areas and pressures of the mirrored elements 2' and 3' are extrapolated from the areas of elements 2 and 6, and 3 and 7 respectively. Likewise the pressure p_M at the corner node is obtained through the pressures of its surrounding elements.

16.8.4   Scheme 5 - Least squares

This scheme is presented (among other places) in Lee *et al.* (1979) [804]. Let us start from the patch of 4 Q_1 elements counting 9 nodes:
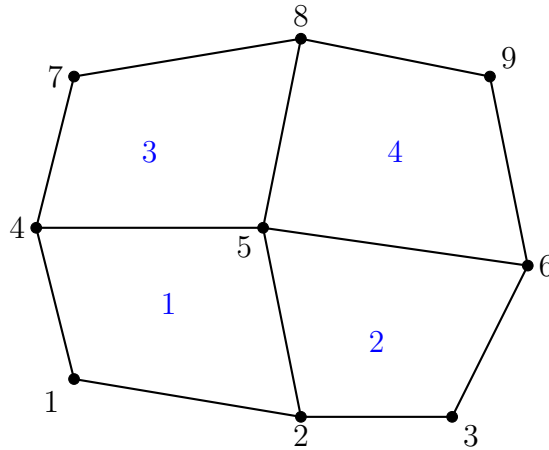
We are looking for a field q living on the nodes. We build the quantity

$$J = \iint_\Omega (q - p)^2 dV$$

where p is the elemental field. To make things clearer we split the integral into the sum of elemental integrals:

$$J = \iint_{\Omega_1} (q(x,y) - p_1)^2 dV + \iint_{\Omega_2} (q(x,y) - p_2)^2 dV + \iint_{\Omega_3} (q(x,y) - p_3)^2 dV + \iint_{\Omega_4} (q(x,y) - p_4)^2 dV$$

Inside each element the field $q(x,y)$ is given by a bilinear interpolation so that:

$$
\begin{aligned}
J \;=\; & \iint_{\Omega_1} (\mathcal{N}_1(x,y)q_1 + \mathcal{N}_2(x,y)q_2 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_4(x,y)q_4 - p_1)^2 dV \\
+ & \iint_{\Omega_2} (\mathcal{N}_2(x,y)q_2 + \mathcal{N}_3(x,y)q_3 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_5(x,y)q_5 - p_2)^2 dV \\
+ & \iint_{\Omega_3} (\mathcal{N}_4(x,y)q_4 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_8(x,y)q_8 + \mathcal{N}_7(x,y)q_7 - p_3)^2 dV \\
+ & \iint_{\Omega_4} (\mathcal{N}_5(x,y)q_5 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_9(x,y)q_9 + \mathcal{N}_8(x,y)q_8 - p_4)^2 dV \qquad (16.24)
\end{aligned}
$$

where the N_i functions are the basis functions (unusually expressed in x, y coordinates). The least square procedure looks for the set of q_i such that

$$\frac{\partial J}{\partial q_i} = 0 \qquad \forall i = 1, ...9$$

and this yields 9 equations/constraints for 9 unknowns.

$$\frac{\partial J}{\partial q_1} = \iint_{\Omega_1} 2(\mathcal{N}_1(x,y)q_1 + \mathcal{N}_2(x,y)q_2 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_4(x,y)q_4 - p_1)\mathcal{N}_1(x,y)dV$$

$$\frac{\partial J}{\partial q_2} = \iint_{\Omega_1} 2(\mathcal{N}_1(x,y)q_1 + \mathcal{N}_2(x,y)q_2 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_4(x,y)q_4 - p_1)\mathcal{N}_2(x,y)dV$$

$$+ \iint_{\Omega_2} 2(\mathcal{N}_2(x,y)q_2 + \mathcal{N}_3(x,y)q_3 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_5(x,y)q_5 - p_2)\mathcal{N}_2(x,y)dV$$

$$\frac{\partial J}{\partial q_3} = \iint_{\Omega_2} 2(\mathcal{N}_2(x,y)q_2 + \mathcal{N}_3(x,y)q_3 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_5(x,y)q_5 - p_2)\mathcal{N}_3(x,y)dV$$

$$\frac{\partial J}{\partial q_4} = \iint_{\Omega_1} 2(\mathcal{N}_1(x,y)q_1 + \mathcal{N}_2(x,y)q_2 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_4(x,y)q_4 - p_1)\mathcal{N}_4(x,y)dV$$

$$+ \iint_{\Omega_3} 2(\mathcal{N}_4(x,y)q_4 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_8(x,y)q_8 + \mathcal{N}_7(x,y)q_7 - p_3)\mathcal{N}_4(x,y)dV$$

$$\frac{\partial J}{\partial q_5} = \iint_{\Omega_1} 2(\mathcal{N}_1(x,y)q_1 + \mathcal{N}_2(x,y)q_2 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_4(x,y)q_4 - p_1)\mathcal{N}_5(x,y)dV$$

$$+ \iint_{\Omega_2} 2(\mathcal{N}_2(x,y)q_2 + \mathcal{N}_3(x,y)q_3 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_5(x,y)q_5 - p_2)\mathcal{N}_5(x,y)dV$$

$$+ \iint_{\Omega_3} 2(\mathcal{N}_4(x,y)q_4 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_8(x,y)q_8 + \mathcal{N}_7(x,y)q_7 - p_3)\mathcal{N}_5(x,y)dV$$

$$+ \iint_{\Omega_4} 2(\mathcal{N}_5(x,y)q_5 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_9(x,y)q_9 + \mathcal{N}_8(x,y)q_8 - p_4)\mathcal{N}_5(x,y)dV$$

$$\frac{\partial J}{\partial q_6} = \iint_{\Omega_2} 2(\mathcal{N}_2(x,y)q_2 + \mathcal{N}_3(x,y)q_3 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_5(x,y)q_5 - p_2)\mathcal{N}_6(x,y)dV$$

$$+ \iint_{\Omega_4} 2(\mathcal{N}_5(x,y)q_5 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_9(x,y)q_9 + \mathcal{N}_8(x,y)q_8 - p_4)\mathcal{N}_6(x,y)dV$$

$$\frac{\partial J}{\partial q_7} = \iint_{\Omega_3} 2(\mathcal{N}_4(x,y)q_4 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_8(x,y)q_8 + \mathcal{N}_7(x,y)q_7 - p_3)\mathcal{N}_7(x,y)dV$$

$$\frac{\partial J}{\partial q_8} = \iint_{\Omega_3} 2(\mathcal{N}_4(x,y)q_4 + \mathcal{N}_5(x,y)q_5 + \mathcal{N}_8(x,y)q_8 + \mathcal{N}_7(x,y)q_7 - p_3)\mathcal{N}_8(x,y)dV$$

$$+ \iint_{\Omega_4} 2(\mathcal{N}_5(x,y)q_5 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_9(x,y)q_9 + \mathcal{N}_8(x,y)q_8 - p_4)\mathcal{N}_8(x,y)dV$$

$$\frac{\partial J}{\partial q_9} = \iint_{\Omega_4} 2(\mathcal{N}_5(x,y)q_5 + \mathcal{N}_6(x,y)q_6 + \mathcal{N}_9(x,y)q_9 + \mathcal{N}_8(x,y)q_8 - p_4)\mathcal{N}_9(x,y)dV \quad (16.25)$$

The factor 2 are removed and the terms $\int p_i N_j$ are known so they end up in the right hand side.

$$\iint_{\Omega_1} (\mathcal{N}_1\mathcal{N}_1 q_1 + \mathcal{N}_1\mathcal{N}_2 q_2 + \mathcal{N}_1\mathcal{N}_5 q_5 + \mathcal{N}_1\mathcal{N}_4 q_4)dV = \iint_{\Omega_1} p_1 N_1 dV$$

$$\iint_{\Omega_1} (\mathcal{N}_2\mathcal{N}_1 q_1 + \mathcal{N}_2\mathcal{N}_2 q_2 + \mathcal{N}_2\mathcal{N}_5 q_5 + \mathcal{N}_2\mathcal{N}_4 q_4)dV$$

$$+ \iint_{\Omega_2} (\mathcal{N}_2\mathcal{N}_2 q_2 + \mathcal{N}_3\mathcal{N}_2 q_3 + \mathcal{N}_6\mathcal{N}_2 q_6 + \mathcal{N}_5\mathcal{N}_2 q_5)dV = \iint_{\Omega_1} p_1 N_2 dV + \iint_{\Omega_2} p_2 \mathcal{N}_2 dV$$

$$\ldots = \ldots$$

$$\iint_{\Omega_4} (\mathcal{N}_9\mathcal{N}_5 q_5 + \mathcal{N}_9\mathcal{N}_6 q_6 + \mathcal{N}_9\mathcal{N}_9 q_9 + \mathcal{N}_9\mathcal{N}_8 q_8)dV = \iint_{\Omega_4} p_4 \mathcal{N}_9 dV \quad (16.26)$$

The mass matrices corresponding to the four elements are

$$
\boldsymbol{M}_1 = \int_{\Omega_1} \begin{pmatrix} \mathcal{N}_1\mathcal{N}_1 & \mathcal{N}_1\mathcal{N}_2 & \mathcal{N}_1\mathcal{N}_5 & \mathcal{N}_1\mathcal{N}_4 \\ \mathcal{N}_2\mathcal{N}_1 & \mathcal{N}_2\mathcal{N}_2 & \mathcal{N}_2\mathcal{N}_5 & \mathcal{N}_2\mathcal{N}_4 \\ \mathcal{N}_5\mathcal{N}_1 & \mathcal{N}_5\mathcal{N}_2 & \mathcal{N}_5\mathcal{N}_5 & \mathcal{N}_5\mathcal{N}_4 \\ \mathcal{N}_4\mathcal{N}_1 & \mathcal{N}_4\mathcal{N}_2 & \mathcal{N}_4\mathcal{N}_5 & \mathcal{N}_4\mathcal{N}_4 \end{pmatrix} dV \qquad
\boldsymbol{M}_2 = \int_{\Omega_2} \begin{pmatrix} \mathcal{N}_2\mathcal{N}_2 & \mathcal{N}_2\mathcal{N}_3 & \mathcal{N}_2\mathcal{N}_6 & \mathcal{N}_2\mathcal{N}_5 \\ \mathcal{N}_3\mathcal{N}_2 & \mathcal{N}_3\mathcal{N}_3 & \mathcal{N}_3\mathcal{N}_6 & \mathcal{N}_3\mathcal{N}_5 \\ \mathcal{N}_6\mathcal{N}_2 & \mathcal{N}_6\mathcal{N}_3 & \mathcal{N}_6\mathcal{N}_6 & \mathcal{N}_6\mathcal{N}_5 \\ \mathcal{N}_5\mathcal{N}_2 & \mathcal{N}_5\mathcal{N}_3 & \mathcal{N}_5\mathcal{N}_6 & \mathcal{N}_5\mathcal{N}_5 \end{pmatrix} dV
$$

$$
\boldsymbol{M}_3 = \int_{\Omega_3} \begin{pmatrix} \mathcal{N}_4\mathcal{N}_4 & \mathcal{N}_4\mathcal{N}_5 & \mathcal{N}_4\mathcal{N}_8 & \mathcal{N}_4\mathcal{N}_7 \\ \mathcal{N}_5\mathcal{N}_4 & \mathcal{N}_5\mathcal{N}_5 & \mathcal{N}_5\mathcal{N}_8 & \mathcal{N}_5\mathcal{N}_7 \\ \mathcal{N}_8\mathcal{N}_4 & \mathcal{N}_8\mathcal{N}_5 & \mathcal{N}_8\mathcal{N}_8 & \mathcal{N}_8\mathcal{N}_7 \\ \mathcal{N}_7\mathcal{N}_4 & \mathcal{N}_7\mathcal{N}_5 & \mathcal{N}_7\mathcal{N}_8 & \mathcal{N}_7\mathcal{N}_7 \end{pmatrix} dV \qquad
\boldsymbol{M}_4 = \int_{\Omega_4} \begin{pmatrix} \mathcal{N}_5\mathcal{N}_5 & \mathcal{N}_5\mathcal{N}_6 & \mathcal{N}_5\mathcal{N}_9 & \mathcal{N}_5\mathcal{N}_8 \\ \mathcal{N}_6\mathcal{N}_5 & \mathcal{N}_6\mathcal{N}_6 & \mathcal{N}_6\mathcal{N}_9 & \mathcal{N}_6\mathcal{N}_8 \\ \mathcal{N}_9\mathcal{N}_5 & \mathcal{N}_9\mathcal{N}_6 & \mathcal{N}_9\mathcal{N}_9 & \mathcal{N}_9\mathcal{N}_8 \\ \mathcal{N}_8\mathcal{N}_5 & \mathcal{N}_8\mathcal{N}_6 & \mathcal{N}_8\mathcal{N}_9 & \mathcal{N}_8\mathcal{N}_8 \end{pmatrix} dV
$$

so that the 9 equations above are actually the result of the assembly process of these four elemental systems:

$$
\left(\iint_{\Omega_e} \vec{\mathcal{N}}^T \vec{\mathcal{N}} dV \right) \cdot \vec{q}_e = \iint_{\Omega_i} \vec{\mathcal{N}}^T p_e dV \qquad e = 1,2,3,4
$$

Also check section 4.5.4 of Glaisner and Tezduyar [493] (1987), in which the authors present a two-step algorithm: 1) pressure is averaged over each element. 2) the nodal values of the pressure are recovered through a least-squares approach.

16.8.5   Scheme 6 - Consistent pressure recovery (penalty formulation)

This is the method presented in Zienkiewicz and Nakazawa [1531] (1982). In the second part of this publication the authors wish to establish a simple and effective numerical method to calculate variables eliminated by the penalisation process. The method involves an additional finite element solution for the nodal pressures using the same finite element basis and numerical quadrature as used for the velocity.

Let us start with[15]:

$$
q = -\lambda \vec{\nabla} \cdot \vec{v}
$$

We are going to treat this equation like any other PDE in the context of the FE method, i.e. we are going to establish its weak form. We assume that the pressure is given inside an element by

$$
q(x,y) = \sum_{i=1}^{m_v} \mathcal{N}_i(x,y) q_i = \vec{\mathcal{N}} \cdot \vec{q}
$$

and the velocity:

$$
\vec{v} = (u,v) \qquad u(x,y) = \sum_{i=1}^{m_v} \mathcal{N}_i(x,y) u_i \qquad v(x,y) = \sum_{i=1}^{m_v} \mathcal{N}_i(x,y) v_i
$$

where the $\mathcal{N}_i$ are the Q_1 basis functions and q_i are the sought after nodal values. We multiply the equation above by a Q_1 basis function $\mathcal{N}_i$ and integrate over the whole domain:

$$
\iint_{\Omega} \mathcal{N}_i(x,y) q(x,y)\, dxdy = -\lambda \iint_{\Omega} \mathcal{N}_i \vec{\nabla} \cdot \vec{v}\, dxdy
$$

As before we now focus on the above expression inside a single element e:

$$
\iint_{\Omega_e} \mathcal{N}_i(x,y) q(x,y)\, dxdy = -\lambda \iint_{\Omega_e} \mathcal{N}_i \vec{\nabla} \cdot \vec{v}\, dxdy
$$

---

[15]I here voluntarily use q instead of p

After $\mathcal{N}_i \to \vec{\mathcal{N}} = (\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4)^T$, the left hand side term becomes:

$$\iint_{\Omega_e} \vec{\mathcal{N}}^T q(x,y) \, dxdy = \iint_{\Omega_e} \vec{\mathcal{N}}^T \vec{\mathcal{N}} \cdot \vec{q} \, dxdy = \left(\underbrace{\iint_{\Omega_e} \vec{\mathcal{N}}^T \vec{\mathcal{N}} dxdy}_{M_e} \right) \cdot \vec{q}$$

where $\boldsymbol{M}_e$ is the elemental mass matrix. We now turn to the right hand side. We have

$$\vec{\nabla} \cdot \vec{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \sum_i \frac{\partial \mathcal{N}_i}{\partial x} u_i + \sum_i \frac{\partial \mathcal{N}_i}{\partial y} v_i$$

We here too define $\vec{V}_e = (u_1, v_1, u_2, v_2, u_3, v_3, u_4, v_4)^T$ so that

$$
\begin{aligned}
&\iint_{\Omega_e} \vec{\mathcal{N}} \vec{\nabla} \cdot \vec{v} \, dV \\
={}& \iint_{\Omega_e} \vec{\mathcal{N}}^T \sum_{i=1}^{m_v} \left(\frac{\partial \mathcal{N}_i}{\partial x} u_i + \frac{\partial \mathcal{N}_i}{\partial y} v_i \right) dV \qquad\qquad (16.27)\\
={}& \iint_{\Omega_e}
\begin{pmatrix}
\mathcal{N}_1 \left(\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x} u_i + \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y} v_i \right) \\
\mathcal{N}_2 \left(\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x} u_i + \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y} v_i \right) \\
\mathcal{N}_3 \left(\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x} u_i + \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y} v_i \right) \\
\mathcal{N}_4 \left(\sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial x} u_i + \sum_{i=1}^{4} \frac{\partial \mathcal{N}_i}{\partial y} v_i \right)
\end{pmatrix} dV \\
={}& \int_{\Omega_e}
\begin{pmatrix}
\mathcal{N}_1 & \mathcal{N}_1 & 0 \\
\mathcal{N}_2 & \mathcal{N}_2 & 0 \\
\mathcal{N}_3 & \mathcal{N}_3 & 0 \\
\mathcal{N}_4 & \mathcal{N}_4 & 0
\end{pmatrix}
\cdot
\begin{pmatrix}
\sum_i \frac{\partial \mathcal{N}_i}{\partial x} u_i \\
\sum_i \frac{\partial \mathcal{N}_i}{\partial y} v_i \\
\sum_i \left(\frac{\partial \mathcal{N}_i}{\partial y} u_i + \frac{\partial \mathcal{N}_i}{\partial x} v_i \right)
\end{pmatrix} dV \\
={}& \int_{\Omega_e}
\underbrace{
\begin{pmatrix}
\mathcal{N}_1 & \mathcal{N}_1 & 0 \\
\mathcal{N}_2 & \mathcal{N}_2 & 0 \\
\mathcal{N}_3 & \mathcal{N}_3 & 0 \\
\mathcal{N}_4 & \mathcal{N}_4 & 0
\end{pmatrix}}_{N}
\cdot
\underbrace{
\begin{pmatrix}
\partial_x \mathcal{N}_1 & 0 & \partial_x \mathcal{N}_2 & 0 & \partial_x \mathcal{N}_3 & 0 & \partial_x \mathcal{N}_4 & 0 \\
0 & \partial_y \mathcal{N}_1 & 0 & \partial_y \mathcal{N}_2 & 0 & \partial_y \mathcal{N}_3 & 0 & \partial_y \mathcal{N}_4 \\
\partial_y \mathcal{N}_1 & \partial_x \mathcal{N}_1 & \partial_y \mathcal{N}_2 & \partial_x \mathcal{N}_2 & \partial_y \mathcal{N}_3 & \partial_x \mathcal{N}_3 & \partial_y \mathcal{N}_4 & \partial_x \mathcal{N}_4
\end{pmatrix}}_{B}
\cdot \vec{V}_e \, dV \\
={}& \left(\int_{\Omega_e} \boldsymbol{N} \cdot \boldsymbol{B} \, dV \right) \cdot \vec{V}_e \\
={}& -\mathbb{G}_e^T \cdot \vec{V}_e \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (16.28)
\end{aligned}
$$

This makes sense since $\mathbb{G}^T$ is the discrete divergence operator. However, it is not very efficient to build $\mathbb{G}_e$ only to multiply it with a vector of already known quantities. In practice we implement Eq. (16.27) which implmentation resembles the buoyancy term of the Stokes equation.

After assembly we arrive at

$$\boldsymbol{M} \cdot \vec{q} = \lambda \mathbb{G}^T \cdot \vec{V} \qquad \text{with} \qquad \mathbb{G}_e = - \int_{\Omega_e} \boldsymbol{N} \cdot \boldsymbol{B} \, dV$$

where $\boldsymbol{M}$ is the global mass matrix, $\vec{q}$ the vector of all nodal pressures, $\mathbb{G}$ the discrete gradient matrix and $\vec{V}$ the (velocity) solution vector. The system can be easily solved since the mass matrix is a friendly matrix. The vector $\vec{q}$ contains the nodal pressure values directly, with no need for a smoothing scheme!

Remark. *Very importantly, the mass matrix $\boldsymbol{M}$ is to be evaluated at the full integration points, while the constraint part (the right hand side of the equation) is to be evaluated at the reduced integration point, i.e. in the middle of the element.*

Remark. *As noted in [1531], it is interesting to note that when linear elements are used and the lumped matrices are used for the $\boldsymbol{M}$ the resulting algebraic equation is identical to the smoothing scheme 1 only if a uniform square finite element mesh is used. In this respect this method is expected to yield different results when elements are not square or even rectangular.*

Remark. *The third column of the matrix $\boldsymbol{N}$ and the last line of the $\boldsymbol{B}$ matrix could be removed altogether. If your code is based on the mixed formulation, then you already have built matrix $\mathbb{G}$ so you can easily re-use this piece of code to compute $\mathbb{G}$ again, this time with a reduced integration quadrature. If you are using the penalty formulation then you need to program all from scratch and then simply do away with these unnecessary terms, or you can direcly build the rhs as $\int_{\Omega_e} \vec{\mathcal{N}}^T p_e$ (assuming you have previously computed the pressure in the middle of each element by means of $p = -\lambda \vec{\nabla} \cdot \vec{v}$).*

Remark. *This scheme is identical to the least square scheme!*

16.8.6 Scheme 7

Same as scheme 6, but with lumped mass matrix.

16.8.7 Scheme 8 - bilinear interpolation

Let us assume that the centers of the four elements make a Q_1 quadrilateral element, as shown on this figure:
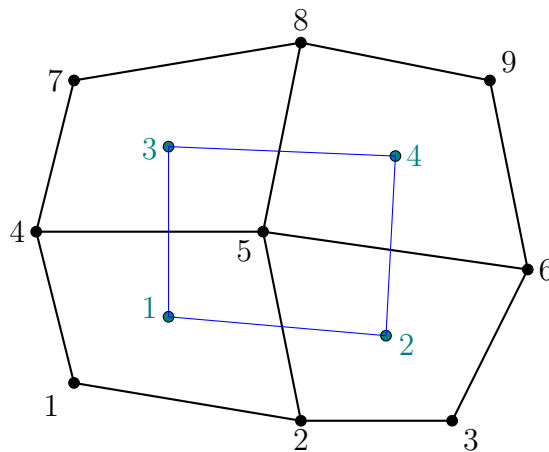
The values at the corners are p_1, p_2, p_3 and p_4. Assuming that the pressure inside this element can be represented by a bilinear field, we have

$$p(x, y) = a + bx + cy + dxy$$

744

where the coefficients will be determined by ensuring that $p(x_i, y_i) = p_i$ for $i = 1, 2, 3, 4$, or:

$$
\begin{aligned}
a + bx_1 + cy_1 + dx_1y_1 &= p_1 & (16.29) \\
a + bx_2 + cy_2 + dx_2y_2 &= p_2 & (16.30) \\
a + bx_3 + cy_3 + dx_3y_3 &= p_3 & (16.31) \\
a + bx_4 + cy_4 + dx_4y_4 &= p_4 & (16.32)
\end{aligned}
$$

i.e.

$$
\begin{pmatrix}
1 & x_1 & y_1 & x_1y_1 \\
1 & x_2 & y_2 & x_2y_2 \\
1 & x_3 & y_3 & x_3y_3 \\
1 & x_4 & y_4 & x_4y_4
\end{pmatrix}
\cdot
\begin{pmatrix}
a \\
b \\
c \\
d
\end{pmatrix}
=
\begin{pmatrix}
p_1 \\
p_2 \\
p_3 \\
p_4
\end{pmatrix}
$$

There remains an issue with nodes which are on the boundaries of the domain. These are of course not 'surrounded' by four pressure values so the above algorithm does not apply directly. However, looking at the above figure, and assuming that node 1 is a lower left corner of a 2D domain, we can use the bilinear interpolation based on elements 1,2,3,4 to extrapolate a nodal pressure value at node 1. The same would apply for nodes 2 and 4 for example.

Remark. *This scheme is not applicable to quadtree-based meshed.*

16.9 The value of the timestep

The chosen time step dt used for time integration is chosen to comply with the Courant-Friedrichs-Lewy condition [22].

$$\delta t = C \min \left(\frac{\min h}{\max |\vec{v}| p}, \frac{h^2}{\kappa} \right) \tag{16.33}$$

where h is a measure of the element diameter, p is the polynomial order of the element, $\kappa = k/\rho C_p$ is the thermal diffusivity and C is the so-called CFL number chosen in $[0, 1[$. $\min h$ is the smallest element diameter in the domain, while $\max |\vec{v}|$ is the maximum velocity (norm) in the domain.

In essence the CFL condition arises when solving hyperbolic PDEs . It limits the time step in many explicit time-marching computer simulations so that the simulation does not produce (too) incorrect results.

This condition is not needed when solving the Stokes equation but it is mandatory when solving the heat transport equation or any kind of advection-diffusion equation. Note that any increase of grid resolution (i.e. h becomes smaller) yields an automatic decrease of the time step value.

16.10   Exporting data to vtk/vtu format

This format seems to be the universally accepted format for 2D and 3D visualisation in Computational Geodynamics (and even CFD ?). Such files can be opened with open source softwares such as Paraview [16], MayaVi [17] or Visit [18].

Unfortunately it is my experience that no simple tutorial exists about how to build such files. There is an official document which describes the vtk format[19] but it delivers the information in a convoluted way[20]. I therefore describe hereafter how fieldstone builds the vtk/vtu files[21].

I hereunder show vtk file corresponding to a 3×2 grid made of linear elements. In this particular example there are:

- 12 nodes and 6 elements

- 1 elemental field (the pressure p)

- 2 nodal fields: 1 scalar (the smoothed pressure q), 1 vector (the velocity field u,v,0)

Note that vtk files are inherently 3D so that even in the case of a 2D simulation the z-coordinate of the points and for instance their z-velocity component must be provided. The file, usually called *solution.vtk* starts with a header:

```
<VTKFile type='UnstructuredGrid' version='0.1' byte_order='BigEndian'>
<UnstructuredGrid>
<Piece NumberOfPoints='12' NumberOfCells='6'>
```

We then proceed to write the node coordinates as follows:

```
<Points>
<DataArray type='Float32' NumberOfComponents='3' Format='ascii'>
0.000000e+00  0.000000e+00  0.000000e+00
3.333333e-01  0.000000e+00  0.000000e+00
6.666667e-01  0.000000e+00  0.000000e+00
1.000000e+00  0.000000e+00  0.000000e+00
0.000000e+00  5.000000e-01  0.000000e+00
3.333333e-01  5.000000e-01  0.000000e+00
6.666667e-01  5.000000e-01  0.000000e+00
1.000000e+00  5.000000e-01  0.000000e+00
0.000000e+00  1.000000e+00  0.000000e+00
3.333333e-01  1.000000e+00  0.000000e+00
6.666667e-01  1.000000e+00  0.000000e+00
1.000000e+00  1.000000e+00  0.000000e+00
</DataArray>
</Points>
```

These are followed by the elemental field(s):

```
<CellData Scalars='scalars'>
<DataArray type='Float32' Name='p' Format='ascii'>
-1.333333e+00
-3.104414e-10
1.333333e+00
-1.333333e+00
```

---

[16]https://www.paraview.org/

[17]https://docs.enthought.com/mayavi/mayavi/

[18]https://wci.llnl.gov/simulation/computer-codes/visit/

[19]https://www.vtk.org/wp-content/uploads/2015/04/file-formats.pdf

[20]I only later realised it was also limited!

[21]I have found the following information about vtk vs. vtu: VTK denotes the simple legacy format file. VTU denotes a Serial Unstructured Grid format information for the XML-based syntax.

```
8.278417e-17
1.333333e+00
</DataArray>
</CellData>
```

Nodal quantities are written next:

```
<PointData Scalars='scalars'>
<DataArray type='Float32' NumberOfComponents='3' Name='velocity' Format='ascii'>
0.000000e+00  0.000000e+00  0.000000e+00
0.000000e+00  0.000000e+00  0.000000e+00
0.000000e+00  0.000000e+00  0.000000e+00
0.000000e+00  0.000000e+00  0.000000e+00
0.000000e+00  0.000000e+00  0.000000e+00
8.888885e-08  -8.278405e-24  0.000000e+00
8.888885e-08  1.655682e-23  0.000000e+00
0.000000e+00  0.000000e+00  0.000000e+00
1.000000e+00  0.000000e+00  0.000000e+00
1.000000e+00  0.000000e+00  0.000000e+00
1.000000e+00  0.000000e+00  0.000000e+00
1.000000e+00  0.000000e+00  0.000000e+00
</DataArray>
<DataArray type='Float32' NumberOfComponents='1' Name='q' Format='ascii'>
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
</DataArray>
</PointData>
```

To these informations we must append 3 more datasets. The first one is the connectivity, the second one is the offsets and the third one is the type. The first one is trivial since the required connectivity array is the same as the one needed for the Finite Elements. The second must be understood as follows: when reading the connectivity information in a linear manner the offset values indicate the beginning of each element (omitting the zero value). The third is simply the type of element as given in the vtk format document (9 corresponds to a generic quadrilateral with an internal numbering consistent with ours - more on this later).

```
<Cells>
<DataArray type='Int32' Name='connectivity' Format='ascii'>
0  1  5  4
1  2  6  5
2  3  7  6
4  5  9  8
5  6  10  9
6  7  11  10
</DataArray>
<DataArray type='Int32' Name='offsets' Format='ascii'>
4
8
12
16
```

```
20
24
</DataArray>
<DataArray type='Int32' Name='types' Format='ascii'>
9
9
9
9
9
9
</DataArray>
</Cells>
```

The file is then closed with

```
</Piece>
</UnstructuredGrid>
</VTKFile>
```

The *solution.vtu*[22] can then be opened with ParaView, MayaVi or Visit and the reader is advised to find tutorials online on how to install and use these softwares. Also check Appendix L.0.1 on how to use ParaView.
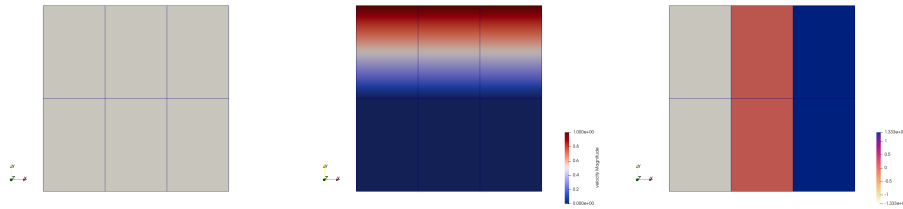
In the same folder `images/vtk` there is the python script makevtu.py[23] which produces 3 different vtu files. The first one *solution1.vtu* is a similar to the one above: an nelx*nely quadrilateral-based mesh in a unit square. The second one (*solution2.vtu*) looks identical when opened in Paraview but it is rather different: each element is exported as its own sub-mesh, so that if the mesh counts nel elements the number of vertices is 4*nel, and not (nelx+1)*(nely+1). As such this file is larger. The icon array is needed to write down the positions of the four vertices of each element but not to write down the connectivity since the first 4 points are making the 1st element, the next four points are making the second element, etc ...

```
vtufile.write("<Points> \n")
vtufile.write("<DataArray type='Float32' NumberOfComponents='3' Format='ascii'> \n")
for iel in range(0,nel):
    if not flag[iel]:
        for k in range(0,m):
            vtufile.write("%10e %10e %10e \n" %(x[icon[k,iel]],y[icon[k,iel]],0.))
vtufile.write("</DataArray>\n")
vtufile.write("</Points> \n")
vtufile.write("<Cells>\n")
vtufile.write("<DataArray type='Int32' Name='connectivity' Format='ascii'> \n")
for iel in range (0,nel_left):
    vtufile.write("%d %d %d %d \n" %(iel*4,iel*4+1,iel*4+2,iel*4+3))
vtufile.write("</DataArray>\n")
...
vtufile.write("</Cells>\n")
```

---

[22]https://raw.githubusercontent.com/cedrict/fieldstone/master/images/vtk/solution.vtu
[23]https://raw.githubusercontent.com/cedrict/fieldstone/master/images/vtk/makevtu.py

This format is rather practical in the case of linear or higher order discontinuous fields. For example, in the case of the $Q_2 \times P_{-1}$ element pair, the pressure is linear inside each element and discontinuous across element edges. One can then assign pressure values at the four vertices of each element.

Finally a third mesh *solution3.vtu* is produced. It is based on the 2nd one, but since elements are now somewhat de-coupled, then one can export only a subset of the mesh. For instance one could not show elements which are too distorted, or below a certain line, or outside a certain volume, etc ... In makevtu.py all elements whose center is inside a circle are flagged and will not be exported into the vtu file:

```
for iel in range(0,nel):
    flag[iel]= (xc[iel]-0.333*Lx)**2+(yc[iel]-0.666*Ly)**2<0.234**2
nel_flagged=np.sum(flag)
nel_left=nel-nel_flagged
```

Once opened in ParaView this is how it looks like:
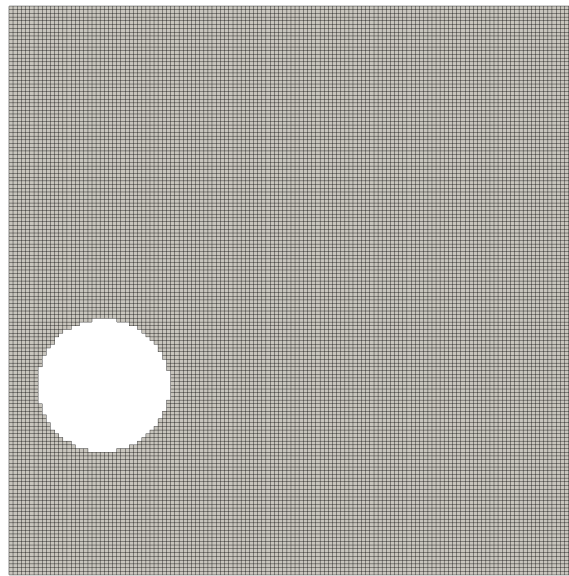
We can find at this address[24] a list of supported cell types (although their internal numbering is not mentioned so enjoy the guess work). Here is a subset of these that can be relevant in our case:

```
// Linear cells
VTK_EMPTY_CELL = 0,
VTK_VERTEX = 1,
VTK_POLY_VERTEX = 2,
VTK_LINE = 3,
VTK_POLY_LINE = 4,
VTK_TRIANGLE = 5,
VTK_TRIANGLE_STRIP = 6,
VTK_POLYGON = 7,
VTK_PIXEL = 8,
VTK_QUAD = 9,
VTK_TETRA = 10,
VTK_VOXEL = 11,
VTK_HEXAHEDRON = 12,
VTK_WEDGE = 13,
```

---

[24]https://vtk.org/doc/nightly/html/vtkCellType_8h_source.html

```
VTK_PYRAMID = 14,
VTK_PENTAGONAL_PRISM = 15,
VTK_HEXAGONAL_PRISM = 16,
// Quadratic, isoparametric cells
VTK_QUADRATIC_EDGE = 21,
VTK_QUADRATIC_TRIANGLE = 22,
VTK_QUADRATIC_QUAD = 23,
VTK_QUADRATIC_POLYGON = 36,
VTK_QUADRATIC_TETRA = 24,
VTK_QUADRATIC_HEXAHEDRON = 25,
VTK_QUADRATIC_WEDGE = 26,
VTK_QUADRATIC_PYRAMID = 27,
VTK_BIQUADRATIC_QUAD = 28,
VTK_TRIQUADRATIC_HEXAHEDRON = 29,
VTK_TRIQUADRATIC_PYRAMID = 37,
VTK_QUADRATIC_LINEAR_QUAD = 30,
VTK_QUADRATIC_LINEAR_WEDGE = 31,
VTK_BIQUADRATIC_QUADRATIC_WEDGE = 32,
VTK_BIQUADRATIC_QUADRATIC_HEXAHEDRON = 33,
VTK_BIQUADRATIC_TRIANGLE = 34,
// Cubic, isoparametric cell
VTK_CUBIC_LINE = 35,
```

After some online searching I found this figure[25]:
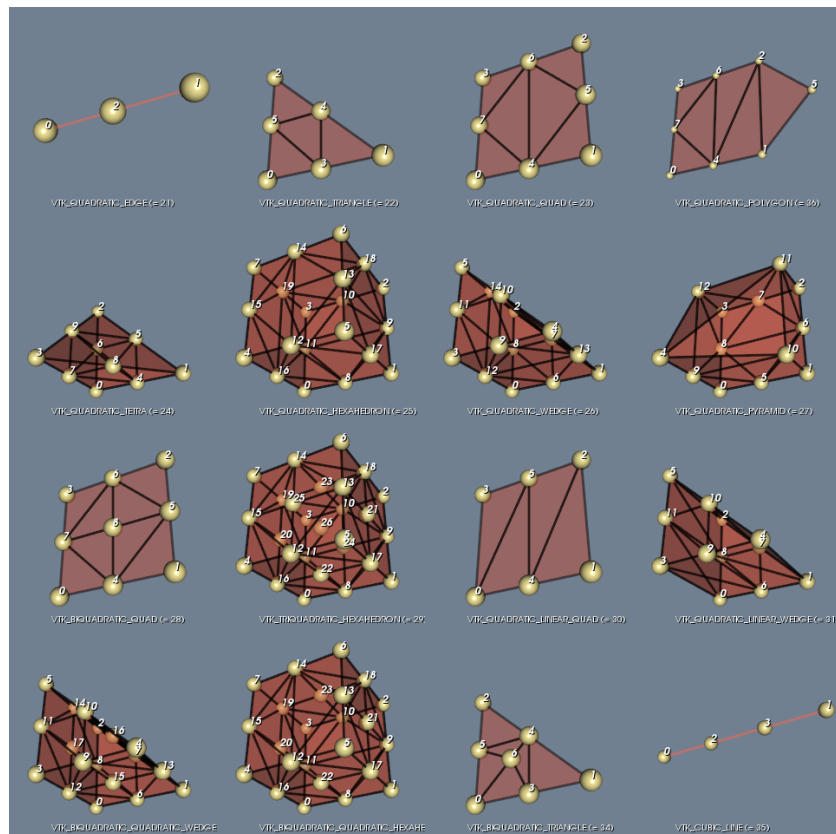
Fortunately these cell types seem to coincide with those in this document [26] mentioned earlier where we find these schematics:

[25]https://examples.vtk.org/site/Cxx/GeometricObjects/IsoparametricCellsDemo/
[26]https://www.vtk.org/wp-content/uploads/2015/04/file-formats.pdf

VTK_VERTEX (=1) VTK_POLY_VERTEX (=2) VTK_LINE (=3)

VTK_POLY_LINE (=4) VTK_TRIANGLE(=5) VTK_TRIANGLE_STRIP (=6)

VTK_POLYGON (=7) VTK_PIXEL (=8) VTK_QUAD (=9)

VTK_TETRA (=10) VTK_VOXEL (=11) VTK_HEXAHEDRON (=12)

VTK_QUADRATIC_EDGE (=21) VTK_QUADRATIC_TRIANGLE (=22) VTK_QUADRATIC_QUAD (=23)

VTK_WEDGE (=13) VTK_PYRAMID (=14)

VTK_QUADRATIC_TETRA (=24) VTK_QUADRATIC_HEXAHEDRON (=25)

752

16.11 Error measurements and convergence rates

What follows is written in the case of a two-dimensional model. Generalisation to 3D is trivial. What follows is mostly borrowed from [1340].

When measuring the order of accuracy of the primitive variables $\vec{v}$ and p, it is standard to report errors in both the L_1 and the L_2 norm. For a scalar quantity Ψ, the L_1 and L_2 norms are computed as

$$\|\Psi\|_1 = \int_V |\Psi| dV \qquad \|\Psi\|_2 = \sqrt{\int_V \Psi^2 dV} \tag{16.34}$$

For a vector quantity $\vec{k} = (k_x, k_y)$ in a two-dimensional space, the L_1 and L_2 norms are defined as:

$$\left\|\vec{k}\right\|_1 = \int_V (|k_x| + |k_y|) dV \qquad \left\|\vec{k}\right\|_2 = \sqrt{\int_V (k_x^2 + k_y^2) dV} \tag{16.35}$$

To compute the respective norms the integrals in the above norms can be approximated by splitting them into their element-wise contributions. The element volume integral can then be easily computed by numerical integration using Gauss-Legendre quadrature.

The respective L_1 and L_2 norms for the pressure error can be evaluated via

$$e_p^h|_1 = \sum_{i=1}^{n_e} \sum_{q=1}^{n_q} |e_p^h(\vec{r}_q)| w_q |J_q| \qquad e_p^h|_2 = \sqrt{\sum_{i=1}^{n_e} \sum_{q=1}^{n_q} |e_p^h(\vec{r}_q)|^2 w_q |J_q|} \tag{16.36}$$

where $e_p^h(\vec{r}_q) = p^h(\vec{r}_q) - p(\vec{r}_q)$ is the pressure error evaluated at the q-th quadrature associated with the ith element. n_e and n_q refer to the number of elements and the number of quadrature points per element. w_q and J_q are the quadrature weight and the Jacobian associated with point q.

The velocity error $e_{\vec{v}}^h$ is evaluated using the following two norms

$$e_{\vec{v}}^h|_1 = \sum_{i=1}^{n_e} \sum_{q=1}^{n_q} [|e_u^h(\vec{r}_q)| + |e_v^h(\vec{r}_q)|] w_q |J_q| \qquad e_{\vec{v}}^h|_2 = \sqrt{\sum_{i=1}^{n_e} \sum_{q=1}^{n_q} [|e_u^h(\boldsymbol{r}_q)|^2 + e_v^h(\boldsymbol{r}_q)|^2] w_q |J_q|} \tag{16.37}$$

where $e_u^h(\vec{r}_q) = u^h(\vec{r}_q) - u(\vec{r}_q)$ and $e_v^h(\vec{r}_q) = v^h(\vec{r}_q) - v(\vec{r}_q)$.

Another norm is very rarely used in the geodynamics literature but is preferred in the Finite Element literature: the H^1 norm. The mathematical basis for this norm and the nature of the $H^1(\Omega)$ Hilbert space is to be found in many FE books [363, 688, 637]. This norm is expressed as follows for a function f such that $f, |\nabla f| \in L^2(\Omega)$ [27]

$$\|f\|_{H^1} = \left(\int_\Omega (|f|^2 + |\nabla f|^2) d\Omega \right)^{1/2} \tag{16.38}$$

We then have

$$e_{\vec{v}}^h|_{H^1} = \left\|\vec{v}^h - \vec{v}\right\|_{H^1} = \sqrt{\sum_{i=1}^{d} \int_\Omega \left[(v_i^h - v_i)^2 + \vec{\nabla}(v_i^h - v_i) \cdot \vec{\nabla}(v_i^h - v_i) \right] d\Omega} \tag{16.39}$$

---

[27] https://en.wikipedia.org/wiki/Sobolev_space

where d is the number of dimensions. Note that sometimes the following semi-norm is used [358, 108]:

$$e_{\vec{v}}^h|_{H^1} = \left\| \vec{v}^h - \vec{v} \right\|_{H^1} = \sqrt{ \sum_{i=1}^{d} \int_{\Omega} \left[\vec{\nabla}(v_i^h - v_i) \cdot \vec{\nabla}(v_i^h - v_i) \right] d\Omega } \tag{16.40}$$

When computing the different error norms for e_p and $e_{\vec{v}}$ for a set of numerical experiments with varying resolution h we expect the error norms to follow the following relationships:

$$e_{\vec{v}}^h|_1 = Ch^{rvL_1} \qquad e_{\vec{v}}^h|_2 = Ch^{rvL_2} \qquad e_{\vec{v}}^h|_{H^1} = Ch^{rvH^1} \tag{16.41}$$

$$e_p^h|_1 = Ch^{rpL_1} \qquad e_p^h|_2 = Ch^{rpL_2} \tag{16.42}$$

where C is a resolution-independent constant and $rpXX$ and $rvXX$ are the convergence rates for pressure and velocity in various norms, respectively. Using linear regression on the logarithm of the respective error norm and the resolution h, one can compute the convergence rates of the numerical solutions.

As mentioned in [358], when finite element solutions converge at the same rates as the interpolants we say that the method is optimal, i.e.:

$$e_{\vec{v}}^h|_{L_2} = \mathcal{O}(h^3) \qquad e_{\vec{v}}^h|_{H^1} = \mathcal{O}(h^2) \qquad e_p^h|_{L_2} = \mathcal{O}(h^2) \tag{16.43}$$

We note that when using discontinuous pressure space (e.g., P_0, P_{-1}), these bounds remain valid even when the viscosity is discontinuous provided that the element boundaries conform to the discontinuity.

About extrapolation

Section contributed by W. Bangerth and part of Thieulot and Bangerth [1345] (2022) but it was ultimately not used.

In a number of numerical benchmarks we want to estimate the error $X_h - X^*$ between a quantity X_h computed from the numerical solution $\vec{v}_h, p_h$ and the corresponding value X computed from the exact solution $\vec{v}, p$. Examples of such quantities X are the root mean square velocity v_{rms}, but it could also be a mass flux across a boundary, an average horizontal velocity at the top boundary, or any other scalar quantity.

If the exact solution is known, then one can of course compute X from it. On the other hand, we would of course like to assess convergence also in cases where the exact solution is not known. In that case, one can compute an *estimate* X^* for X by way of *extrapolation*. To this end, we make the assumption that asymptotically, X_h converges to X at a fixed (but unknown) rate r, so that

$$e_h = |X_h - X| \approx Ch^r. \tag{16.44}$$

Here, X, C and r are all unknown constants to be determined, although we are not really interested in C. We can evaluate X_h from the numerical solution on successively refined meshes with mesh sizes h, $h/2$, and $h/4$. Then, in addition to (16.44) we also have

$$e_{h/2} = |X_{h/2} - X| \approx C\left(\frac{h}{2}\right)^r, \tag{16.45}$$

$$e_{h/4} = |X_{h/4} - X| \approx C\left(\frac{h}{4}\right)^r. \tag{16.46}$$

Taking ratios of equations (16.44)–(16.46), and replacing the unknown X by an *estimate* X^*, we then arrive at the following equation:

$$\frac{|X_h - X^\star|}{|X_{h/2} - X^\star|} = \frac{|X_{h/2} - X^\star|}{|X_{h/4} - X^\star|} = 2^r.$$

If one assumes that X_h converges to X uniformly either from above or below (rather than oscillate around X), then this equation allows us to solve for X^* and r:

$$(X_h - X^\star)(X_{h/4} - X^\star) = (X_{h/2} - X^\star)(X_{h/2} - X^\star)$$

$$X_h X_{h/4} - X^\star X_{h/4} - X_h X^\star + (X^\star)^2 = X_{h/2}^2 - 2X^\star X_{h/2} + (X^\star)^2$$

$$X_h X_{h/4} - X^\star X_{h/4} - X_h X^\star = X_{h/2}^2 - 2X^\star X_{h/2}$$

$$X_h X_{h/4} - X_{h/2}^2 = -2X^\star X_{h/2} + X^\star X_{h/4} + X_h X^\star$$

$$X_h X_{h/4} - X_{h/2}^2 = X^\star(-2X_{h/2} + X_{h/4} + X_h)$$

and finally:

$$X^\star = \frac{X_h X_{h/4} - X_{h/2}^2}{X_h - 2X_{h/2} + X_{h/4}}, \qquad r = \log_2 \frac{X_{h/2} - X^\star}{X_{h/4} - X^\star}.$$

In the determination of r, we could also have used X_h and $X_{h/2}$, but using $X_{h/2}$ and $X_{h/4}$ is generally more reliable because the higher order terms we have omitted in (16.44) are less visible on finer meshes.

In some cases, however, halving the mesh size multiple times is not really tractable (memory problem, or cpu time). Let us now start again from

$$e_h = |X_h - X| \approx C h^r. \tag{16.47}$$

and assume that we run two other models at a resolution αh and βh, such that $1 > \alpha > \beta > 0$. In the example above we of course had $\alpha = 1/2$ and $\beta = 1/4$. Then we have

$$e_{\alpha h} = |X_{\alpha h} - X| \approx C(\alpha h)^r, \tag{16.48}$$
$$e_{\beta h} = |X_{\beta h} - X| \approx C(\beta h)^r. \tag{16.49}$$

which leads to

$$\frac{|X_h - X^\star|}{|X_{\alpha h} - X^\star|} = \frac{Ch^r}{C(\alpha h)^r} = (1/\alpha)^r \qquad \text{and} \qquad \frac{|X_{\alpha h} - X^\star|}{|X_{\beta h} - X^\star|} = \frac{C(\alpha h)^r}{C(\beta h)^r} = (\alpha/\beta)^r$$

In order for both to be equal we must have

$$(1/\alpha)^r = (\alpha/\beta)^r \qquad \Rightarrow \qquad 1/\alpha = \alpha/\beta \qquad \Rightarrow \qquad \beta = \alpha^2$$

So of course if $\alpha = 1/2$ then $\beta = 1/4$, but now we can also take $\alpha = 3/4$ and then $\beta = 9/16$. Etc ...

In the end, this approach might not be that useful since the mesh sizes would then be $h, 3h/4, 9h/16, 27h/64$ which may be hard to achieve in practice.

16.12 The initial temperature field

📙 Relevant Literature:

- Thermal gradients in the continental crust [236]

- Simple analytical approximation to the temperature structure in subduction zones [401]

- Thermal structure of subduction zone back arcs [315]

- Thermal Structure of Oceanic Lithosphere [1137]

- thermal structure of a subducting plate with finite length [631]

Single layer with imposed temperature b.c.

Let us take a single layer of material characterised by a heat capacity C_p, a heat conductivity k and a heat production term H.

The Heat transport equation writes

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot (k \vec{\nabla} T) + \rho H \tag{16.50}$$

At steady state and in the absence of a velocity field, assuming that the material properties to be independent of time and space, and assuming that there is no heat production ($H = 0$), this equation simplifies to

$$\Delta T = 0 \tag{16.51}$$

Assuming the layer to be parallel to the x-axis, the temperature is $T(x, y) = T(y) = \alpha T + \beta$. In order to specify the constants α and β, we need two constraints.

At the bottom of the layer $y = y_b$ a temperature T_b is prescribed while a temperature T_t is prescribed at the top with $y = y_t$. This ultimately yields a temperature field in the layer given by

$$T(y) = \frac{T_t - T_b}{y_t - y_b}(y - y_b) + T_b$$

If now the heat production coefficient is not zero, the differential equation reads

$$k \Delta T + H = 0 \tag{16.52}$$

The temperature field is then expected to be of the form

$$T(y) = -\frac{H}{2k} y^2 + \alpha y + \beta \tag{16.53}$$

Supplied again with the same boundary conditions, this leads to

$$\beta = T_b + \frac{H}{2k} y_b^2 - \alpha y_b$$

ie,
$$T(y) = -\frac{H}{2k}(y^2 - y_b^2) + \alpha(y - y_b) + T_b$$

and finally
$$\alpha = \frac{T_t - T_b}{y_t - y_b} + \frac{H}{2k}(y_b + y_t)$$

or,
$$T(y) = -\frac{H}{2k}(y^2 - y_b^2) + \left(\frac{T_t - T_b}{y_t - y_b} + \frac{H}{2k}(y_b + y_t)\right)(y - y_b) + T_b$$

Taking $H = 0$ in this equation obviously yields the temperature field obtained previously. Taking $k = 2.25$, $T_t = 0C$, $T_b = 550C$, $y_t = 660km$, $y_b = 630km$ yields the following temperature profiles and heat fluxes when the heat production H varies:
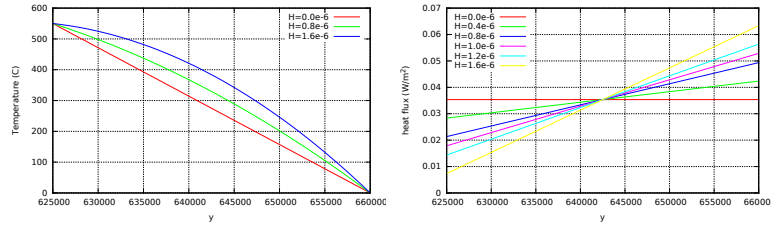
Looking at the values at the top, which are somewhat estimated to be about $55 - 65 mW/m^2$ [677, table 8.6], one sees that value $H = 0.8e - 6$ yields a very acceptable heat flux. Looking at the bottom, the heat flux is then about $0.03 W/m^2$ which is somewhat problematic since the heat flux at the Moho is reported to be somewhere between 10 and 20 mW/m^2 in [677, table 7.1].

Single layer with imposed heat flux b.c.

Let us now assume that heat fluxes are imposed at the top and bottom of the layer:
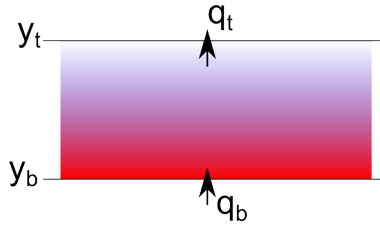
We start again from the ODE
$$k\Delta T + H = 0$$

but only integrate it once:
$$k\frac{dT}{dy} + Hy + \alpha = 0$$

At the bottom $q = k(dT/dy)|_{y=y_b} = q_b$ and at the top $q = k(dT/dy)|_{y=y_t} = q_t$ so that

to finish

Single layer with imposed heat flux and temperature b.c.
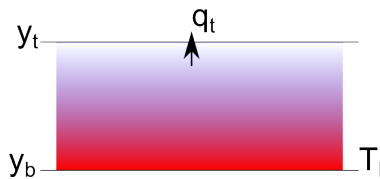
to finish

Half cooling space

TODO.

Plate model

[905]

McKenzie slab

When doing thermo-mechanical modelling, the initial temperature field in the domain is of prime importance. This is especially true for the temperature in the slab for subduction modelling as its rheological behaviour is strongly temperature-dependent. One could easily design a simple geometrical initial field but it is unlikely to be close to the field of a slowly subducting slab at an angle in a hot mantle.

McKenzie [901] derived such approximate initial field from the steady-state energy equation in two dimensions:

$$\rho C_p \vec{v} \cdot \vec{\nabla} T = k \vec{\nabla}^2 T \tag{16.54}$$

We denote by T_l the temperature at the base of the lithosphere and l its thickness (i.e. the thickness of the slab).

Assuming $\vec{v} = (v_x, 0)$ yields

$$\rho C_p v_x \frac{\partial T}{\partial x} = k \frac{\partial^2 T}{\partial x^2}$$

and substitution of $T' = T/T_l$, $x' = x/l$ and $z' = z/l \in [0, 1]$ in this equation leads to

$$\rho C_p v_x \frac{T_l}{l} \frac{\partial T'}{\partial x'} = k \frac{T_l}{l^2} \left(\frac{\partial^2 T'}{\partial x'^2} + \frac{\partial^2 T'}{\partial z'^2} \right)$$

or

$$\frac{\rho C_p v_x l}{k} \frac{\partial T'}{\partial x'} = \frac{\partial^2 T'}{\partial x'^2} + \frac{\partial^2 T'}{\partial z'^2}$$

and finally (see Eq. 2.3 of [901]):

$$\frac{\partial^2 T'}{\partial x'^2} - 2R \frac{\partial T'}{\partial x'} + \frac{\partial^2 T'}{\partial z'^2} = 0$$

where R is the thermal Reynolds number

$$R = \frac{\rho C_p v_x l}{2k}$$

The general solution to this PDE with $T' = 1$ on the top, left and right boundary is

$$T'(x', z') = 1 + \sum_n C_n \exp \left[\left(R - (R^2 + n^2 \pi^2)^{1/2} \right) x' \right] \sin(n\pi z')$$

We now must make an assumption about the temperature on the left boundary ($x' = 0$), which is the temperature of the lithosphere. For simplicity McKenzie assumes that $T'(x' = 0, z') = 1 - z'$ so that $C_n = 2(-1)^n / n\pi$ and finally

$$T'(x', z') = 1 + 2 \sum_n \frac{(-1)^n}{n\pi} \exp \left[\left(R - (R^2 + n^2 \pi^2)^{1/2} \right) x' \right] \sin(n\pi z') \tag{16.55}$$

Let us build a simple temperature model for a 250km × 50km slab, with $\rho = 3000$, $C_p = 1250$, $k = 3$. The python code is available in `images/mckenzie/mckenzie1.py`.
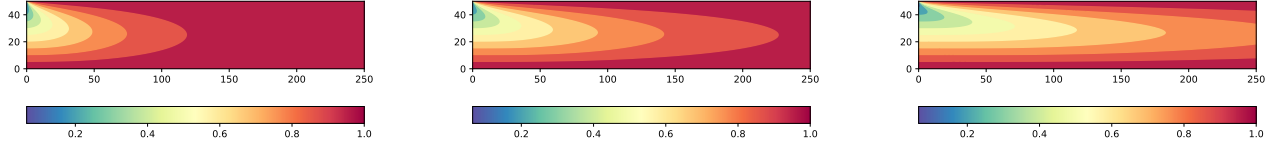
Left to right: Dimensionless temperature T' in a 250km × 50km slab for $v_x = 0.5, 1, 2$cm/year

We logically recover the fact that the slower the slab penetrates the mantle the more temperature diffusion dominates over temperature advection. For $v = 0.5$cm/year we see that that the slab assumes a constant temperature $T' = 1$ at all depthes $0 \leq z' \leq 1$ for $x' \geq 125$km.

Note that this field is a steady-state field, valid for a constant density, heat conductivity and heat capacity, zero heat production, that it implies that the velocity is constant and that the lithosphere temperature is linear.

One can also embed the slab in a more realistic context, a subduction zone, involving a subducting lithosphere, an over-riding plate and a mantle. The domain is 1000km × 250km. The mantle temperature is set to 1300 °C. The slab dip can be varied and so can the velocity. The python code is available in `images/mckenzie/mckenzie2.py`.
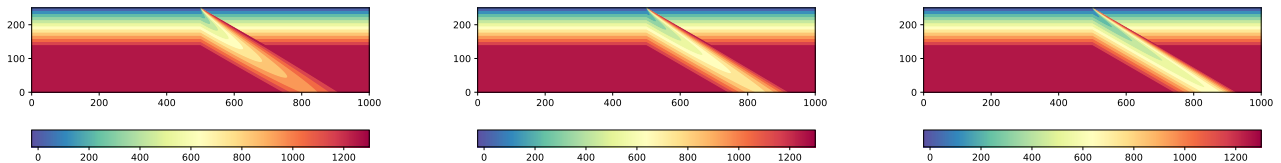
Left to right: temperature T for $v_x = 0.5, 1, 2$cm/year and $\phi = 30°$.
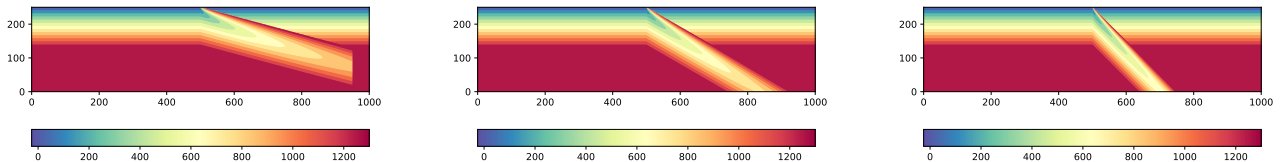
Left to right: temperature T for $v_x = 1$cm/year and $\phi = 15, 30, 45°$.

Initial temperature for global mantle convection models

This is a difficult topic, and Gottschaldt *et al.* [505] list a few issues or facts to take into account:

759

- Frequent impacts may have determined the heat structure of the outer layers (Arrhenius and Lepland 2000), leading to an early thermally stable stratification.

- A global magma ocean (Solomatov 2000) or several large scale melting events (Kleine *et al.* . 2004) are also conceivable.

- Fractional crystallisation and subsequent overturn has the potential to result in compositionally or thermally stable layering, too (Elkins-Tanton *et al.* 2003; Zaranek and Parmentier 2004)

16.13 The consistent boundary flux (CBF)

The Consistent Boundary Flux technique was devised to alleviate the problem of the accuracy of primary variables derivatives (mainly velocity and temperature) on boundaries. These derivatives are important since they are needed to compute the heat flux (and therefore the Nusselt number) or dynamic topography and geoid.

The idea was first introduced in Mizukami (1986) [937] and later used in geodynamics in Zhong *et al.* (1993) [1518]. It was finally implemented in the CITCOMS code [1521, 950] and more recently in the ASPECT code (dynamic topography postprocessor). Note that the CBF should be seen as a post-processor step as it does not alter the primary variables values.

The CBF method is implemented and used in STONE 27. It is also discussed but not explicitly named in Reddy's book [1118, p309]. Also see Larock & Herrmann (1976) [790], Gresho *et al.* (1987) [516], Marshall *et al.* [885].

16.13.1 The CBF applied to the Stokes equation

We start from the strong form:

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{b} = \vec{0} \tag{16.56}$$

and then write the weak form on an element e:

$$\int_{\Omega_e} N_i^\gamma \vec{\nabla} \cdot \boldsymbol{\sigma} \, dV + \int_{\Omega_e} N_i^\gamma \vec{b} \, dV = \vec{0} \tag{16.57}$$

We then use the two equations:

$$\vec{\nabla} \cdot (N \boldsymbol{\sigma}) = N \vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{\nabla} N \cdot \boldsymbol{\sigma} \qquad \text{(chain rule)}$$

$$\int_\Omega (\vec{\nabla} \cdot \boldsymbol{\sigma}) \, dV = \int_\Gamma \boldsymbol{\sigma} \cdot \vec{n} \, dS \qquad \text{(divergence theorem)}$$

and integrate by parts in order to obtain:

$$\int_\Gamma N_i^\gamma \boldsymbol{\sigma} \cdot \vec{n} \, dS - \int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \boldsymbol{\sigma} \, dV + \int_{\Omega_e} N_i^\gamma \vec{b} \, dV = \vec{0} \tag{16.58}$$

and since the traction vector $\vec{t}$ is given by $\vec{t} = \boldsymbol{\sigma} \cdot \vec{n}$ we have:

$$\int_{\Gamma_e} N_i^\gamma \vec{t} \, dS = \int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \boldsymbol{\sigma} \, dV - \int_{\Omega_e} N_i^\gamma \vec{b} \, dV \tag{16.59}$$

The core idea of the method lies in considering the traction vector as an unknown living on the nodes on the boundary, and assuming we have already solved the Stokes equation and therefore have obtained the velocity and pressure.

Finally, since the traction vector can be expressed as a function of the velocity basis functions on the edge i.e.

$$\vec{t} = \sum_{i=1}^m N_i^\gamma \vec{t_i}$$

the left hand term yields an edge (1D) mass matrix $\mathbb{M}'$ (see Section C).

Remark. *In* STONE *27 an alternative to equation 16.59 is used. Although somewhat inefficient, the elemental matrices $\mathbb{K}$ and $\mathbb{G}$ and the corresponding body force rhs are built and the rhs of the traction equation is computed as follows:*

$$\mathbb{M}' \cdot \vec{\mathcal{T}} = -\mathbb{K} \cdot \vec{\mathcal{V}} - \mathbb{G} \cdot \vec{\mathcal{P}} + \vec{f}$$

where $\vec{\mathcal{T}}$ is the vector of assembled tractions which we want to compute and $\vec{\mathcal{V}}$ and $\vec{\mathcal{T}}$ are the solutions of the Stokes problem.

Remark. *The assembled mass matrix is tri-diagonal and can be easily solved with a Conjugate Gradient method.*

Remark. *With a trapezoidal integration rule (i.e. Gauss-Lobatto - see Section 7.2.7) the matrix can even be diagonalised and the resulting matrix is simply diagonal, which results in a very cheap solve as mentioned in Zhong et al. (1993) [1518].*

Some implementation details for the Stokes equation

What follows is relevant for STONE 27 which relies on Q_1 shape functions for the velocity. Let us start with a small example, a 3x2 element FE grid:
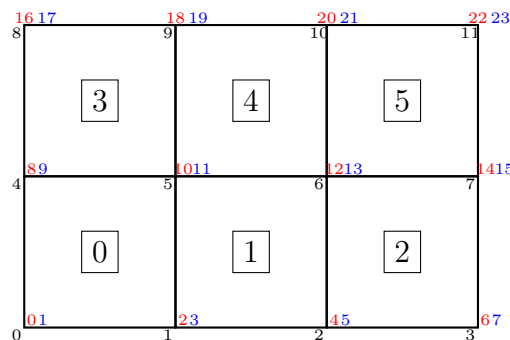
Red color corresponds to the dofs in the x direction, blue color indicates a dof in the y direction.

We have nnp=12, nel=6, NfemV=24. Let us assume that free slip boundary conditions are applied. The boundary conditions `fix_bc` array is then:

bc_fix=[T T T T T T T T T T T T T T T T T T T T T T T T]

Note that since corners belong to two edges, we effectively prescribed no-slip boundary conditions on those.

why does array contain only T??

We wish to compute the tractions on the boundaries, and more precisely for the dofs for which a Dirichlet velocity boundary condition has been prescribed. The number of (traction) unknowns NfemTr is then the number of T in the `bc_fix` array. In our specific case, we wave NfemTr= . This means that we need for each targeted dof to be able to find its identity/number between 0 and NfemTr-1. We therefore create the array `bc_nb` which is filled as follows:

finish

bc_nb=[T T T T T T T T T T T T T T T T T T T T]

This translates as follows in the code:

762

```
NfemTr=np.sum(bc_fix)
bc_nb=np.zeros(NfemV,dtype=np.int32)
counter=0
for i in range(0,NfemV):
    if (bc_fix[i]):
        bc_nb[i]=counter
        counter+=1
```

The algorithm is then as follows

A Prepare two arrays to store the matrix M_{cbf} and its right hand side rhs_{cbf}

B Loop over all elements

C For each element touching a boundary, compute the residual vector $R_{el} = -f_{el} + \mathbb{K}_{el}\mathcal{V}_{el} + \mathbb{G}_{el}\mathcal{P}_{el}$

D Loop over the four edges of the element using the connectivity array

E For each edge loop over the number of degrees of freedom (2 in 2D)

F For each edge assess whether the dofs on both ends are target dofs.

G If so, compute the mass matrix M_{edge} for this edge

H extract the 2 values off the element residual vector and assemble these in rhs_{cbf}

I Assemble M_{edge} into NfemTrxNfemTr matrix using bc_nb

```
M_cbf = np.zeros((NfemTr,NfemTr),np.float64)          # A
rhs_cbf = np.zeros(NfemTr,np.float64)

for iel in range(0,nel):                              # B

    ... compute elemental residual ...                # C

  #boundary 0-1                                        # D
  for i in range(0,ndofV):                             # E
        idof0=2*icon[0,iel]+i
        idof1=2*icon[1,iel]+i
        if (bc_fix[idof0] and bc_fix[idof1]):          # F
            idofTr0=bc_nb[idof0]
            idofTr1=bc_nb[idof1]
            rhs_cbf[idofTr0]+=res_el[0+i]              # H
            rhs_cbf[idofTr1]+=res_el[2+i]
            M_cbf[idofTr0,idofTr0]+=M_edge[0,0]        #
            M_cbf[idofTr0,idofTr1]+=M_edge[0,1]        # I
            M_cbf[idofTr1,idofTr0]+=M_edge[1,0]        #
            M_cbf[idofTr1,idofTr1]+=M_edge[1,1]        #

  #boundary 1-2                                         #[D]

  ...

  #boundary 2-3                                         #[D]

  ...

  #boundary 3-0                                         #[D]

  ...
```

16.13.2 The CBF applied to the heat transport equation

We start from the strong form of the heat transfer equation (without the source terms for simplicity):

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = \vec{\nabla} \cdot k \vec{\nabla} T$$

The weak form then writes:

$$\int_\Omega \mathcal{N}^\theta \rho C_p \frac{\partial T}{\partial t} dV + \int_\Omega \mathcal{N}^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T dV = \int_\Omega \mathcal{N}^\theta \vec{\nabla} \cdot k \vec{\nabla} T dV$$

Using once again integration by parts and divergence theorem:

$$\int_\Omega \mathcal{N}^\theta \rho C_p \frac{\partial T}{\partial t} dV + \int_\Omega \mathcal{N}^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T dV = \int_\Gamma \mathcal{N}^\theta k \vec{\nabla} T \cdot \vec{n} d\Gamma - \int_\Omega \vec{\nabla} \mathcal{N}^\theta \cdot k \vec{\nabla} T dV$$

On the boundary we are interested in the heat flux $\vec{q} = -k \vec{\nabla} T$

$$\int_\Omega \mathcal{N}^\theta \rho C_p \frac{\partial T}{\partial t} dV + \int_\Omega \mathcal{N}^\theta \rho C_p \vec{v} \cdot \vec{\nabla} T dV = -\int_\Gamma \mathcal{N}^\theta \boldsymbol{q} \cdot \boldsymbol{n} d\Gamma - \int_\Omega \vec{\nabla} \mathcal{N}^\theta \cdot k \vec{\nabla} T dV$$

or,

$$\int_\Gamma \mathcal{N}^\theta \boldsymbol{q} \cdot \boldsymbol{n} d\Gamma = -\int_\Omega \mathcal{N}^\theta \rho C_p \frac{\partial T}{\partial t} dV - \int_\Omega \mathcal{N}^\theta \rho C_p \boldsymbol{v} \cdot \vec{\nabla} T dV - \int_\Omega \vec{\nabla} \mathcal{N}^\theta \cdot k \vec{\nabla} T dV$$

Considering the normal heat flux $q_n = \vec{q} \cdot \vec{n}$ as an unknown living on the nodes on the boundary,

$$q_n = \sum_{i=1}^{2} q_{n|i} N_i$$

so that the left hand term becomes a mass matrix for the basis functions living on the boundary. We have already covered the right hand side terms when building the FE system to solve the heat transport equation, so that in the end

$$\mathbb{M}' \cdot \vec{\mathcal{Q}}_n = -\mathbb{M} \cdot \frac{\partial \boldsymbol{T}}{\partial t} - K_a \cdot \boldsymbol{T} - K_d \cdot \boldsymbol{T}$$

where $\vec{\mathcal{Q}}_n$ is the assembled vector of normal heat flux components. Note that in all terms the assembly only takes place over the elements along the boundary.

Note that the resulting matrix is symmetric.

16.14 Static condensation

The idea behind static condensation is quite simple: in some cases, there are dofs belonging to an element which only belong to that element. For instance, the so-called MINI element ($P_1^+ \times P_1$) showcases a bubble function in the middle (see section **??**). In the following, $\vec{\mathcal{V}}^\star$ corresponds to the list of such dofs inside an element. The discretised Stokes equations on any element looks like:

$$
\begin{pmatrix}
\mathbb{K} & L & \mathbb{G} \\
L^T & \mathbb{K}^\star & H \\
\mathbb{G}^T & H^T & 0
\end{pmatrix}_e
\begin{pmatrix}
\vec{\mathcal{V}} \\
\vec{\mathcal{V}}^\star \\
\vec{\mathcal{P}}
\end{pmatrix}_e
=
\begin{pmatrix}
\vec{f} \\
\vec{f}^\star \\
\vec{h}
\end{pmatrix}_e
\tag{16.60}
$$

This is only a re-writing of the elemental Stokes matrix where the matrix $\mathbb{K}$ has been split in four parts. Note that the matrix $\mathbb{K}^\star$ is diagonal.

This can also be re-written in non-matrix form:

$$
\mathbb{K} \cdot \vec{\mathcal{V}} + L \cdot \vec{\mathcal{V}}^\star + \mathbb{G} \cdot \vec{\mathcal{P}} = \vec{f} \tag{16.61}
$$
$$
L^T V + K^\star \cdot \vec{\mathcal{V}}^\star + H \cdot \vec{\mathcal{P}} = \vec{f}^\star \tag{16.62}
$$
$$
\mathbb{G}^T \cdot \vec{\mathcal{V}} + H^T \vec{\mathcal{V}}^\star = \vec{h} \tag{16.63}
$$

The $\vec{\mathcal{V}}^\star$ in the second equation can be isolated:

$$
\vec{\mathcal{V}}^\star = \mathbb{K}^{-\star} \cdot (\vec{f}^\star - L^T \cdot \vec{\mathcal{V}} - H \cdot \vec{\mathcal{P}})
$$

and inserted in the first and third equations:

$$
\mathbb{K} \cdot \vec{\mathcal{V}} + L \left[\mathbb{K}^{-\star}(\vec{f}^\star - L^T \cdot \vec{\mathcal{V}} - H \cdot \vec{\mathcal{P}}) \right] + \mathbb{G} \cdot \vec{\mathcal{P}} = \vec{f} \tag{16.64}
$$
$$
\mathbb{G}^T \cdot \vec{\mathcal{V}} + H^T \left[\mathbb{K}^{-\star}(\vec{f}^\star - L^T \cdot \vec{\mathcal{V}} - H \cdot \vec{\mathcal{P}}) \right] = \vec{h} \tag{16.65}
$$

or,

$$
(\mathbb{K} - L \cdot \mathbb{K}^{-\star} \cdot L^T) \cdot \vec{\mathcal{V}} + (G - L \cdot \mathbb{K}^{-\star} \cdot H) \cdot \vec{\mathcal{P}} = \vec{f} - L \cdot \mathbb{K}^{-\star} \cdot \vec{f}^\star \tag{16.66}
$$
$$
(G^T - H^T \cdot \mathbb{K}^{-\star} \cdot L^T) \cdot \vec{\mathcal{V}} - (H^T \cdot \mathbb{K}^{-\star} \cdot H) \cdot \vec{\mathcal{P}} = \vec{h} - H^T \cdot \mathbb{K}^{-\star} \cdot \vec{f}^\star \tag{16.67}
$$

i.e.

$$
\underline{\mathbb{K}} \cdot \vec{\mathcal{V}} + \underline{\mathbb{G}} \cdot \vec{\mathcal{P}} = \underline{\vec{f}} \tag{16.68}
$$
$$
\underline{\mathbb{G}}^T \cdot \vec{\mathcal{V}} - \underline{\mathbb{C}} \cdot \vec{\mathcal{P}} = \underline{\vec{h}} \tag{16.69}
$$

with

$$
\underline{\mathbb{K}} = K - L \cdot \mathbb{K}^{-\star} \cdot L^T \tag{16.70}
$$
$$
\underline{\mathbb{G}} = G - L \cdot \mathbb{K}^{-\star} \cdot H \tag{16.71}
$$
$$
\underline{\mathbb{C}} = H^T \cdot \mathbb{K}^{-\star} \cdot H \tag{16.72}
$$
$$
\underline{\vec{f}} = \vec{f} - L \cdot \mathbb{K}^{-\star} \cdot \vec{f}^\star \tag{16.73}
$$
$$
\underline{\vec{h}} = \vec{h} - H^T \cdot \mathbb{K}^{-\star} \cdot \vec{f}^\star \tag{16.74}
$$

Note that $\underline{\mathbb{K}}$ is symmetric, and so is the Stokes matrix.

For instance, in the case of the MINI element, the dofs corresponding to the bubble could be eliminated at the elemental level, which would make the Stokes matrix smaller (see book by Braess [136]). However, it is then important to note that static condensation introduces a pressure-pressure term which was not there in the original formulation. This is also presented in the appendix of Karabelas *et al.* [711].

16.15 Measuring incompressibility

The velocity divergence error integrated over the whole element is given by

$$e_{div} = \int_{\Omega} (\vec{\nabla} \cdot \vec{v}^h - \underbrace{\vec{\nabla} \cdot \vec{v}}_{=0}) \, d\Omega = \int_{\Omega} (\vec{\nabla} \cdot \vec{v}^h) \, d\Omega \tag{16.75}$$

where Γ_e is the boundary of element e and $\vec{n}$ is the unit outward normal of Γ_e.

Furthermore, one can show that [358]:

$$e_{div} = \int_{\Gamma_e} \vec{v}^h \cdot \vec{n} \, d\Gamma$$

The reason is as follows and is called the divergence theorem[28]: suppose a volume V subset of $\mathbb{R}^d$ which is compact and has a piecewise smooth boundary S, and if $\vec{F}$ is a continuously differentiable vector field then

$$\int_{V} (\vec{\nabla} \cdot \vec{F}) \, dV = \int_{S} (\vec{F} \cdot \vec{n}) \, dS$$

The left side is a volume integral while the right side is a surface integral. Note that sometimes the notation $d\vec{S} = \vec{n} \, dS$ is used so that $\vec{F} \cdot \vec{n} \, dS = \vec{F} \cdot d\vec{S}$.

The average velocity divergence over an element can be defined as

$$< \vec{\nabla} \cdot \vec{v} >_e = \frac{1}{V_e} \int_{\Omega_e} (\vec{\nabla} \cdot \vec{v}) \, d\Omega = \frac{1}{V_e} \int_{\Gamma_e} \vec{v} \cdot \vec{n} \, d\Gamma$$

Note that for elements using discontinuous pressures we shall recover a zero divergence element per element (local mass conservation) while for continuous pressure elements the mass conservation is guaranteed only globally (i.e. over the whole domain), see section 3.13.2 of [517].

Note that one could instead compute $< |\vec{\nabla} \cdot \vec{v}| >_e$. Either volume or surface integral can be computed by means of an appropriate Gauss-Legendre quadrature algorithm.

---

[28]https://en.wikipedia.org/wiki/Divergence_theorem

16.16 Parallel or not?

Rationale

Let us assume that we want ro run a simulation of the whole Earth mantle with a constant resolution of 5km. The volume of the mantle is

$$V_{mantle} = \frac{4}{3}\pi(R_{out}^3 - R_{in}^3) \simeq 10^{12}km^3$$

while the volume of an element is $V_e = 125km^3$ (this is only an average since the tesselation of a hollow sphere with hexahedra yields elements which are not all similar [1344]). Consequently, the number of cells needed to discretise the mantle is

$$N_{el} = \frac{V_{mantle}}{V_e} \simeq 8 \times 10^9$$

We know that the matrix size is approx. 4 times the number of elements in 3D:

$$N \simeq 25 \times 10^9$$

Using between 9 and 125 particles per element (a very conservative number), the total number of particles is then

$$N_{particles} \geq 10^{10}$$

The unescapable conclusion is that high-resolution 3D calculations have a very large memory footprint and require extremely long computational times.

The only way to overcome this problem is by resorting to using supercomputers with many processors and large memory capacities.

The idea behind parallel programming is to have each processor carry out only a subset of the total number of operations required. In order to reduce the memory footprint on each processor, only a subset of the computational mesh is known by each: one speaks then of domain decomposition [1364].

An example of such a large parallel calculation of 3D convection with domain decomposition in a spherical shell can be found in [776]:

a)   b)
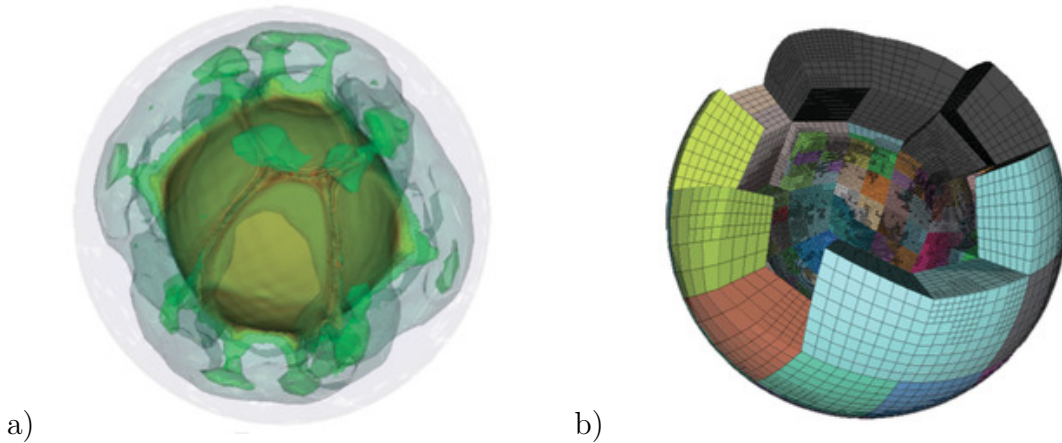
a)Isocontours of the temperature field; b) Partitioning of the domain onto 512 proc. The mesh counts 1,424,176 cells. The solution has approximately 54 million unknowns (39 million vel., 1.7 million press., and 13 million temp.)

Basic approaches

In the past, many applications implemented the idea below on the left using 1D domain decomposition (also known as "slab decomposition"). In the following left figure, a 3D domain is arbitrarily chosen to be decomposed in Y and X directions. It can be seen that in state (a), any computations in the X-Z planes can be done in local memories while data along a Y mesh-line is distributed.
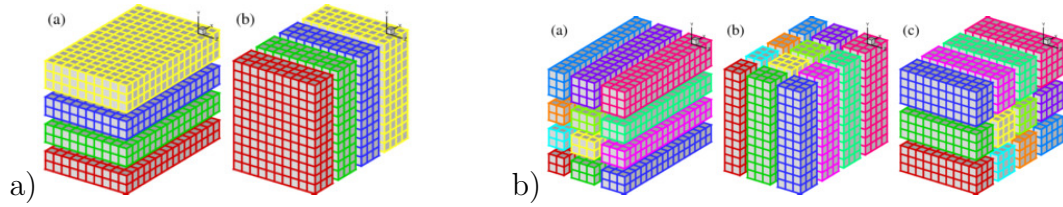
Left: 1D domain decomposition example using 4 processors: (a) decomposed in Y direction; (b) decomposed in X direction. Right: 2D domain decomposition example using a 4x3 processor grid.

A $1D$ decomposition, while quite simple, has some limitations, especially for large-scale simulations. Given a cubic mesh of size N^3, one obvious constraint is that the maximum number of processors Nproc that can be used in a 1D decomposition is N as each slab has to contain at least one plane [29] For a cubic mesh with $1000^3 = 10^9$ points, the constraint is $Nproc \leq 1000$. This is a serious limitation as most supercomputers today have more than 10,000 cores and some have more than 100,000. Large applications are also likely to hit the memory limit when each processor handles too much workload.

Load balancing

In computing, load balancing distributes workloads across multiple computing resources, such as computers, a computer cluster, network links, central processing units or disk drives. Load balancing aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any single resource.

More concretely, the use of many processors in the case where the mesh is unstructured and highly irregular raises the question of how the partitioning is done so that each processor gets a similar workload, thereby minimising latency and optimising cpu time.

This is a difficult problem which is often addressed by means of graph partitioners, and which is made more complicated by the use of AMR.

A typical strategy goes as follows. One can prove that **space filling curves** can be generated for any regular subdivision of a square (see dashed line in the following figure).
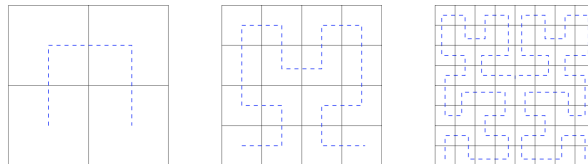
It now the refined grid is the one shown on the following figure, one can use the curve to 'write' all elements in a line and then cut this curve in Nproc chunks (in this case 4, hence four colours).

---
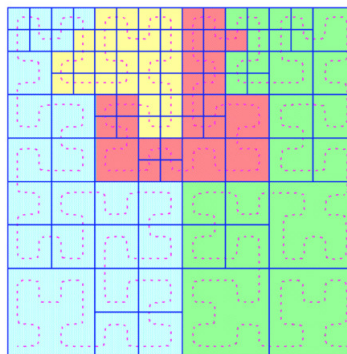
[29]This domain decomposition approach is the one carried out in the FANTOM code [1343] and this limitation was often encountered when using grids such as 160x160x23, thereby limiting the number of processors to about 100 processors.

The mesh counts in total 100 elements and each domain counts 25 elements. This decomposition looks very appealing but another aspect must be taken in account: communication.

Ideally, one wishes to carry out the domain decomposition in a simple-enough manner so that it does not require too much work, and also in such a way that the surface between between all the domain is minimised, since it is related to the amount of communication across processors.

The following figures showcase examples of domain decomposition projected onto a complex 3D geometry.
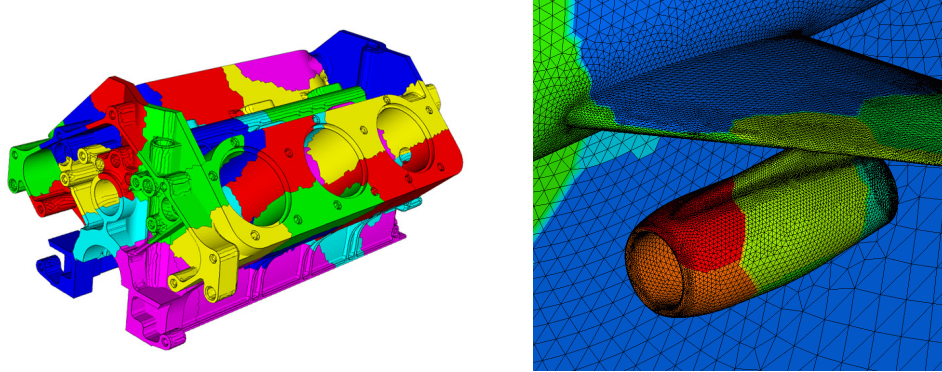
Right: Domain decomposition for parallel processing of a wing-body-nacelle-pylon configuration. Each colour corresponds to a different processor.

Strong scaling vs weak scaling
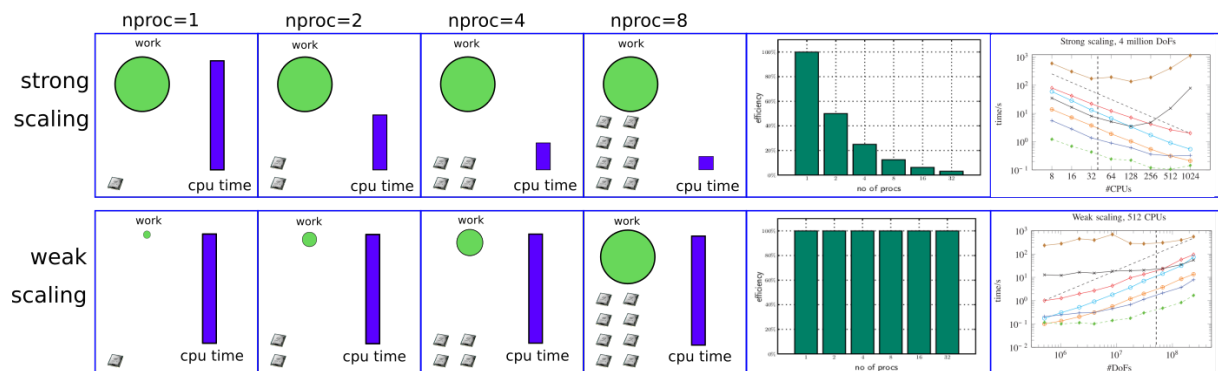
16.17 Corner flow

The mantle wedge comprised between the downgoing slab and the overriding plate has been extensively studied since very important geodynamical processes take place in it or right above it (slab dehydration and water transport, melting, over-riding plate deformation, vulcanism, ...).

To first approximation one can approach the problem and simplify it greatly by assuming that both plates kinematic behaviour are independent of what happens in the wedge, that the wedge geometry does not change over time, that the problem is essentially 2D, and that the mantle extends very far away from the actual wedge (plates are infinite).

Under such assumptions, it is possible to derive an analytical solution for incompressible Stokes flow in the wedge as documented at p. 224 in Batchelor [56].

Literature: [1367]

FIND refs. check new version of Vol7 theoretical geophys

A corner flow setup is shown hereunder:
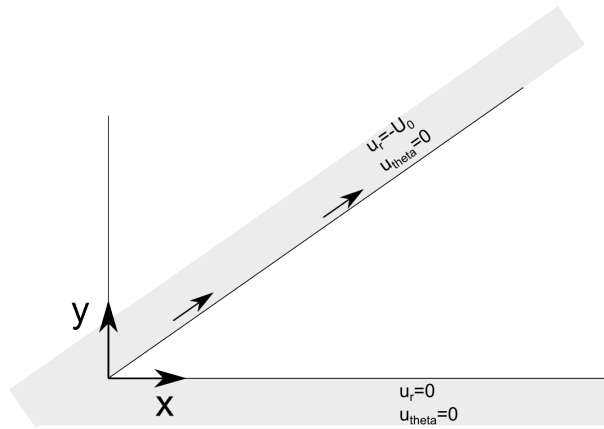
The solution to this problem is arrived at by means of the stream function Φ, defined as $u = -\partial\Phi/\partial y$ and $v = \partial\Phi/partial x$, so that we automatically have $\vec{\nabla} \cdot \vec{v} = 0$. As shown in Section **??**, the stream function Φ is then the solution to the biharmonic equation

$$\vec{\nabla}^2\vec{\nabla}^2\Phi = \vec{\nabla}^4\Phi = 0$$

Considering the geometry of the problem has plates of infinite extent with constant relative velocity, the solution for velocity everywhere is expected to be independent of r. This means the equation is separable and we will use a solution of the form

$$\Phi(r,\theta) = R(r)f(\theta)$$

However, given the infinite extent of the domain, the velocity is expected to be independent of r, so we postulate $R(r) = r$ (look at the relationship between velocity components and stream function), or:

$$\Phi(r,\theta) = rf(\theta)$$

and we then have to solve

$$\Delta\left(\frac{1}{r}(f + f'')\right) = \frac{1}{r^3}(f + 2f'' + f'''') = 0.$$

The solution of this equation for f is:

$$
\begin{aligned}
f(\theta) &= A\sin\theta + B\cos\theta + C\theta\sin\theta + D\theta\cos\theta \\
f'(\theta) &= A\cos\theta - B\sin\theta + C(\sin\theta + \theta\cos\theta) + D(\cos\theta - \theta\sin\theta)
\end{aligned}
$$

with

$$\nu_r = \frac{1}{r}\frac{\partial \Phi}{\partial \theta} = f'(\theta)$$

$$\nu_\theta = -\frac{\partial \Phi}{\partial r} = -f(\theta)$$

A, B, C and D are four constants to be determined by means of the boundary conditions which are as follows:

$$\begin{aligned}
\nu_r(\theta = 0) &= 0 \\
\nu_\theta(\theta = 0) &= 0 \\
\nu_r(\theta = \theta_0) &= -U_0 \\
\nu_\theta(\theta = \theta_0) &= 0
\end{aligned}$$

or,

$$\begin{aligned}
f'(0) = A + D &= 0 && (16.76) \\
f(0) = B &= 0 && (16.77) \\
f'(\theta_0) &= -U_0 && (16.78) \\
f(\theta_0) &= 0 && (16.79)
\end{aligned}$$

From the second equation it is trivial to see that $B = 0$, so that:

$$f(\theta) = A\sin\theta + C\theta\sin\theta + D\theta\cos\theta$$

$$f'(\theta) = A\cos\theta + C(\sin\theta + \theta\cos\theta) + D(\cos\theta - \theta\sin\theta)$$

From the first one we obtain $D = -A$ so that

$$f(\theta) = A(\sin\theta - \theta\cos\theta) + C\theta\sin\theta$$

$$f'(\theta) = A(\theta\sin\theta) + C(\sin\theta + \theta\cos\theta)$$

The last two boundary conditions yield:

$$0 = A(\sin\theta_0 - \theta_0\cos\theta_0) + C\theta_0\sin\theta_0$$

$$-U_0 = A(\theta_0\sin\theta_0) + C(\sin\theta_0 + \theta_0\cos\theta_0)$$

or,

$$A = -U_0\frac{\theta_0\sin\theta_0}{\theta_0^2 - \sin^2\theta_0} \qquad C = U_0\frac{\sin\theta_0 - \theta_0\cos\theta_0}{\theta_0^2 - \sin^2\theta_0}$$

Finally:

$$(A, B, C, D) = (-\theta_0\sin\theta_0, 0, \sin\theta_0 - \theta_0\cos\theta_0, \theta_0\sin\theta_0)\frac{U_0}{\theta_0^2 - \sin^2\theta_0}$$

We have

$$\begin{aligned}
\boldsymbol{e}_r &= \cos\theta\boldsymbol{e}_x + \sin\theta\boldsymbol{e}_y && (16.80) \\
\boldsymbol{e}_\theta &= -\sin\theta\boldsymbol{e}_x + \cos\theta\boldsymbol{e}_y && (16.81)
\end{aligned}$$

so that the velocity field can be expressed in Cartesian coordinates:

$$\begin{aligned}
\boldsymbol{\nu} &= \nu_r\boldsymbol{e}_r + \nu_\theta\boldsymbol{e}_\theta \\
&= \nu_r(\cos\theta\boldsymbol{u}_x + \sin\theta\boldsymbol{e}_y) + \nu_\theta(-\sin\theta\boldsymbol{u}_x + \cos\theta\boldsymbol{e}_y) \\
&= (\nu_r\cos\theta - \nu_\theta\sin\theta)\boldsymbol{e}_x + (\nu_r\sin\theta + \nu_\theta\cos\theta)\boldsymbol{e}_y && (16.82)
\end{aligned}$$

⬙ Relevant Literature: Ribe [1131] present a simple model for the mantle flow induced by back arc spreading behind a subduction zone.

16.18   Surface processes

In 1D - simple nonlinear diffusion a la Burov & Cloetingh (1997)

This approach comes from Burov and Cloetingh [197] (1997). The tectonic-scale transport equations describe long term changes in topography $h(x, y, t)$ as a result of simultaneous short- and long-range mass transport processes [64, 765].

The short-range surface processes are represented by cumulative effects of hillslope processes (soil creep, rainsplash, slides) that remove material from uplifted areas down to the valleys. It is then assumed that the horizontal material flux $\vec{q}_s$ is related to local slope $\vec{\nabla}h$ by $\vec{q}_s = -K_s\vec{\nabla}h$ where K_s is the effective diffusivity. Assumption of conservation of mass volume leads to the linear diffusion equation for erosion:

$$\frac{\partial h}{\partial t} = K_s \Delta h$$

This equation can be solved with constant-elevation (fixed h value) boundary conditions simulating local base levels of erosion.

Note that is practice the coefficient K_s might depend on slope and curvature, i.e.

$$\frac{\partial h}{\partial t} = K_s(x, y, h, \vec{\nabla}h) \Delta h$$

Following [503], Burov & Cloetingh use an empirical non linear expression $K_s = k_s(x)(\vec{\nabla}h)^n$.

In 1D - not so simple, a la Andrès-Martinez *et al.* (2019)

This approach comes from Andrés-Martínez, Pérez-Gussinyé, Armitage, and Morgan [25] (2019). The change in surface elevation rate due to surface processes is equal to the divergence of the sediment flux (assuming there is no density difference between the bedrock and sediment and ignoring the effects of compaction):

$$\frac{\partial h}{\partial t} = -\frac{\partial q_s}{\partial x}$$

where h is the topography, t is the time, q_s represents the sediment flux, and x is the horizontal coordinate.

The next step consists in a formulation for the sediment flux. Still following [25], in the subaerial environment, it is possible to define the sediment transport flux q_s in terms of the water flux q_w as

$$q_s = -(K + cq_w^n)\frac{\partial h}{\partial x}$$

where K is the slope diffusivity, c is the transport coefficient, and $n \geq 1$ is the power law that defines the type of relationship between the sediment transport and the water flux (Simpson & Schlunegger, 2003; Smith & Bretherton, 1972).

get these papers

This model accounts for hillslope diffusion processes where the topography will tend to a dispersive diffusion (Culling, 1960) and fluvial transport processes that result in concentrative diffusion due to water run off (Graf, 1984). For a simple parameterization we choose a linear relationship between sediment transport and water flux ($n = 1$).

The water flux can be related to the water discharge/effective rainfall α as

$$\frac{\partial}{\partial x}(\vec{n}q_w) = -\alpha$$

where $\vec{n}$ is a unit vector directed down the surface gradient (Smith & Bretherton, 1972). By assuming a constant α and integrating equation (12) over the surface in the downstream direction, we obtain

$$q_w = \alpha x_d$$

where x_d is the downstream distance from the drainage divide. By substituting equations (11) and (13) into (10) we obtain the 1-D sediment mass conservation equation for combined hillslope and discharge-dependent fluvial transport

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left((K + k\alpha x_d)\frac{\partial h}{\partial x}\right)$$

where the downstream distance x_d is calculated at each time step as the distance from the topographic highs to the valley floors. Because q_w is dependent on the length of the drainage, the model mimics 1-D landscapes similar to river profiles in which fluvial processes are dominant.

Examples in the literature
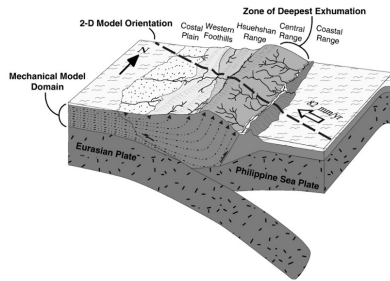
Fig. 4. Schematic block model of Taiwan represented as an orogenic wedge. Material is added to the orogen through frontal accretion and underplating in the west and removed through erosion on the surface. The general west-to-east transport and the high erosion rates lead to an asymmetric pattern of exhumation with the highest rates in the ECR. Our two-dimensional numerical model is oriented parallel to the convergence direction, and the mechanical domain of the model is parameterized to represent the deforming crust above the main detachment.
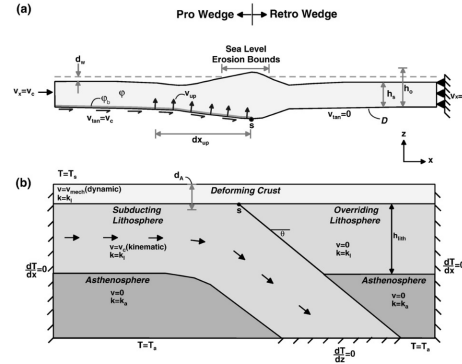
Fig. 5. Model domains, boundary conditions and parameters (Table 3) for the coupled mechanical and thermal model used in this study. (a) The domain of the mechanical model represents the deforming crust above the main detachment. The detachment is simulated using a region of lower internal angle of friction ($\varphi_b < \varphi$) along the base of the model. Deformation within this domain is calculated using temperature dependent viscous and

Application to Taiwan. Taken from Fuller *et al.* (2006) [451]

16.19 Geometric multigrid

The following is mostly borrowed from the Wikipedia page on multigrid methods[30].

There are many types of (geometric) multigrid algorithms, but the common features are that a hierarchy of grids is considered. The important steps are:

- *Smoothing*: reducing high frequency errors, for example using a few iterations of the Gauss-Seidel method.

- *Residual Computation*: computing residual error after the smoothing operation(s).

- *Restriction*: downsampling the residual error to a coarser grid.

- *Interpolation or prolongation*: interpolating a correction computed on a coarser grid into a finer grid.

- *Correction*: Adding prolongated coarser grid solution onto the finer grid.

There are many choices of multigrid methods with varying trade-offs between speed of solving a single iteration and the rate of convergence with said iteration. The 3 main types are V-Cycle, F-Cycle, and W-Cycle.

Any geometric multigrid cycle iteration is performed on a hierarchy of grids and hence it can be coded using recursion. Since the function calls itself with smaller sized (coarser) parameters, the coarsest grid is where the recursion stops.

Note that the ratio of the number of nodes between two consecutive levels has to be constant between all the levels. Often powers of 2 are used (especially if the grids are based on quad/octrees) but it is not a requirement.

Image from `http://web.utk.edu/~wfeng1/research.html`
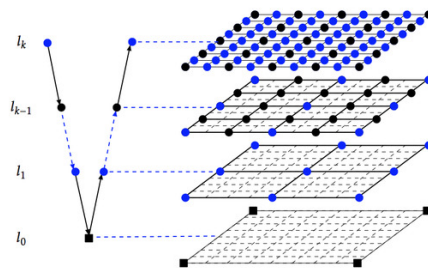
What follows is a pseudo-code example of a recursive V-Cycle Multigrid for solving the Poisson equation ($\nabla^2 \phi = f$) on a uniform grid of spacing h:

```
function phi = V_Cycle(phi,f,h)
% Pre-Smoothing
phi = smoothing(phi,f,h);
% Compute Residual Errors
r = residual(phi,f,h);
% Restriction
rhs = restriction(r);
eps = zeros(size(rhs));
% stop recursion at smallest grid size
if smallest_grid_size_is_achieved
```

---

[30]`https://en.wikipedia.org/wiki/Multigrid_method`

```
    eps = smoothing(eps,rhs,2*h);
else
    eps = V_Cycle(eps,rhs,2*h);
end
% Prolongation and Correction
phi = phi + prolongation(eps);
% Post-Smoothing
phi = smoothing(phi,f,h);
end
```

A multigrid method with an intentionally reduced tolerance can be used as an efficient precon-ditioner for an external iterative solver. The solution may still be obtained in $\mathcal{O}(N)$ time as well as in the case where the multigrid method is used as a solver. Multigrid preconditioning is used in practice even for linear systems, typically with one cycle per iteration.
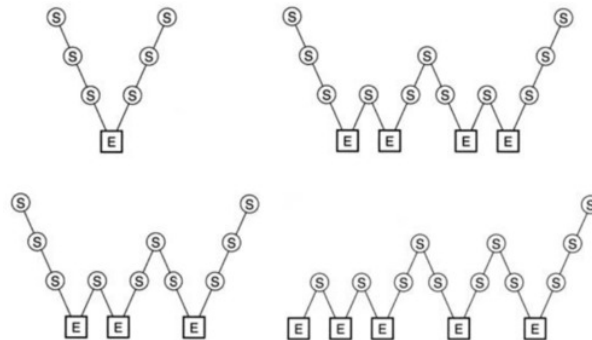
Taken from [659]: Different types of multigrid cycle with four grid levels: (top left) V-cycle, (top right) W-cycle, (bottom left) F-cycle and (bottom right) full multigrid. 'S' denotes smoothing while 'E' denotes exact coarse-grid solution.

Check Kaus BEcker syllabus!

📚 Relevant Literature: [1063, 838, 1374, 659, 483, 894, 850, 1375, 961, 1482, 1451, 1332, 1484, 278] Book [166]

- ACuTEMan: A multigrid-based mantle convection simulation code and its optimization to the Earth Simulator, Kameyama (2005) [705]

16.20 Algebraic multigrid

Relevant Literature: [985][1003]

16.21 Computing depth

In the case of a perfectly rectangular, cylindrical or spherical domain, computing the depth of any given point inside the domain is trivial. However, when the free surface becomes somewhat distorted, the concept of depth needs to be refined. What follows is an attempt at bringing clarity as to how to compute depth in all cases.

The depth $d(\boldsymbol{r})$ satisfies the equation:

$$\frac{\boldsymbol{g}}{|\boldsymbol{g}|} \cdot \boldsymbol{\nabla} d = 1$$

with $d = 0$ at the surface.

This is a form of steady-state advection equation (the time derivative is zero, there is no diffusion, nor any source term).

Given the boundary conditions, one could solve this equation over the whole domain.

Note that in the case of a cartesian box, $\boldsymbol{g} = -g\boldsymbol{u}_z$, we need to solve

$$-\frac{\partial}{\partial z} d = 1$$

For a flat top surface at $d(z = L_z) = 0$ so that in the end

$$d(z) = L_z - z$$

16.22 The Geoid

What is the geoid?

There is an infinity of equipotential surfaces of the gravitational potential U. However, there is a particular surface on the Earth that is "easy" to locate: the mean sea level. This is a somewhat arbitrary choice but it makes sense because the oceans are made of water (!): the surface of a fluid in equilibrium must follow an equipotential.
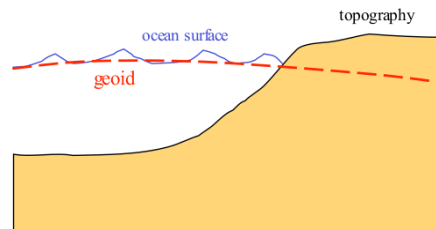
The geoid is usually defined in two ways:

- it is the particular equipotential surface that coincides with the mean sea level (easy to define in the oceans -assuming no currents, waves,... - but harder on land since it is not the topographic surface).

- A gravitational equipotential surface. This means that everywhere at sea level experiences the same value of gravity potential, so there is no tendency for water to flow downhill since all points in the vicinity have the same value of gravity potential, pointed toward the center of the earth.
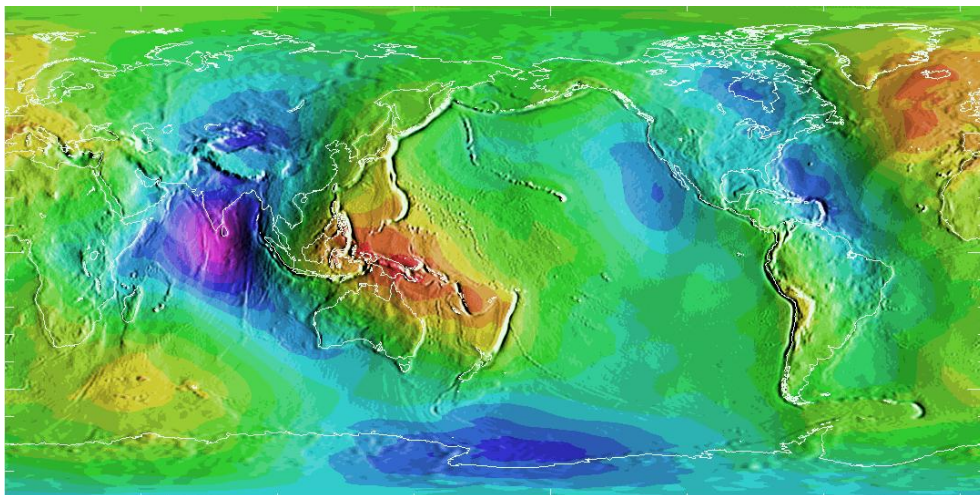
Data Max value: 85.4 meters, east of New Guinea. Data Min value:-107.0 meters, south of India. This image shows 15'x15' geoid undulations covering the planet Earth from the NIMA/GSFC WGS-84 EGM96 15' Geoid Height File. The undulations refer to the differences from the WGS-84(G873) reference ellipsoid. Map and description from National Geodetic Survey [31]

From Wikipedia: The geoid surface is irregular, unlike the reference ellipsoid (which is a mathematical idealized representation of the physical Earth), but is considerably smoother than Earth's physical surface. Although the physical Earth has excursions of +8,848 m (Mount Everest) and -11,034 m (Marianas Trench), the geoid's deviation from an ellipsoid ranges from +85 m (Iceland) to -106 m (southern India), less than 200 metre total.

---

[31] https://www.usna.edu/Users/oceano/pguth/md_help/geology_course/geoid.htm

1. Ocean 2. Reference ellipsoid 3. Local plumb line 4. Continent 5. Geoid

Taken from `https://en.wikipedia.org/wiki/Geoid`
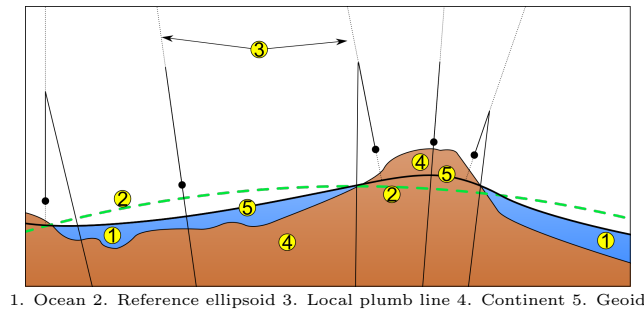
the (reference) ellipsoid

First evidence that the Earth is round Erathostene (275-195 B.C.)

First hypothesis that the Earth is flattened at the poles: Newton

First measurement of the Earth's flattening at the poles: Clairaut (1736) and Bouguer (1743)

The shape of the Earth can be mathematically represented as an ellipsoid defined by:

- Semi-major axis = equatorial radius = a

- Semi-minor axis = polar radius = c

- Flattening (the relationship between equatorial and polar radius): $f = (a - c)/a$

- Eccentricity: $e^2 = 2f - f^2$

Many different reference ellipsoids have been defined and are in use. We define the *reference ellipsoid* = the ellipsoid that best fits the geoid. It is totally arbitrary, but practical. The most common reference ellipsoid is the WGS-84 one[32]:
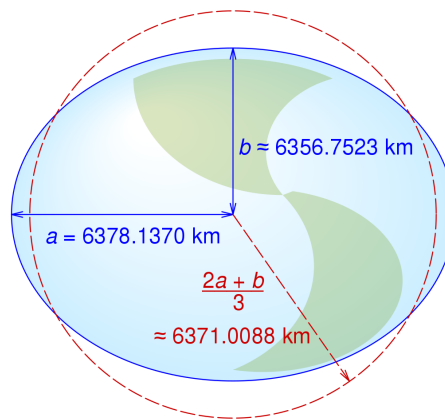
Taken from `https://en.wikipedia.org/wiki/Reference_ellipsoid`

Geoid undulations = differences, in meters, between the geoid reference ellipsoid (= geoid "height"). To clarify:

- Geoid = the equipotential surface of the Earth's gravity field that best fits (in a least squares sense) the mean sea level. The gravitational potential is constant on the geoid (by definition) but the gravitational acceleration is not!

- Reference Ellipsoid = the ellipsoid that best fits the geoid

- Geoid = the (actual) figure of the Earth

- Ellipsoid = the (theoretical) shape of the Earth

---

[32]`https://confluence.qps.nl/qinsy/latest/en/world-geodetic-system-1984-wgs84-182618391.html`

How to compute it?

From Liu and Zhong [849] (2016): " The geoid is computed by ϕ/g, where ϕ is the surface gravitational potential anomaly and can be solved from the Poisson equation, $\nabla^2\phi = -4\pi\mathcal{G}\delta\rho$ where $\mathcal{G}$ is the gravitational constant, and $\delta\rho$ includes both density variations in the mantle [...] and those associated with dynamic topographies at the surface and CMB. Dynamic topographies are determined from solving [the Stokes] equations under free-slip boundary conditions at the surface and CMB."

Interesting modelling
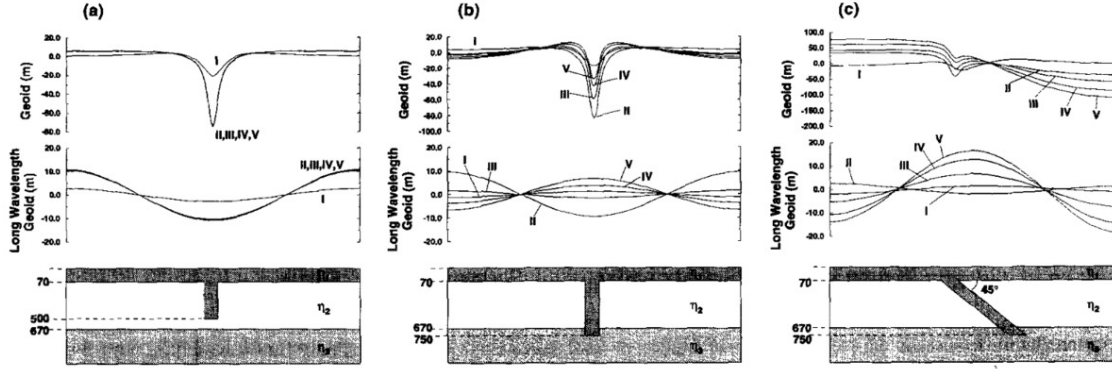
Idealized 2D slab calculations for each viscosity model: geoid and geoid filtered to pass only the longest wavelengths ($\sim$ 4000 km). (a) Cold slab extends to 500 km depth in the upper mantle, (b) Slab extends to 750 km so that it is partly supported by the high viscosity lower mantle at 670 km. (c) Slab tilted at 45° to the vertical extending to the top of the lower mantle. Taken from [953]

16.23 Mixing and stirring, the Lyapunov time/exponent

Many approaches are taken in the literature when it comes to studying mixing/stirring in fluids, and in our case the mantle.

For example, Samuel and Tosi [1180] (2012) measure the convective stirring efficiency using two Lagrangian methods: the first determines the mixing time associated with different wavelengths of heterogeneity following the approach of Ferrachat and Ricard [417] (2001) The second determines the value of the maximum Finite Time Lyapunov Exponents (FTLE) as described in Farnetani and Samuel [410] (2003), and measures the rate at which heterogeneities are stretched by mantle motions.

In the review of van der Wiel, van Hinsbergen, Thieulot, and Spakman [1401] by Nicolas Coltice, his review reads:

> " The first studies of mantle mixing date back to the 1980's and one of the difficulty was to deal with scales. So 2 approaches were taken: one dealing explicitly with scales, such as the authors do here with configurational entropy. It was in relationship with the observations of heterogeneities at all scales (marble cake like at first, more comprehensive nowadays with mid-ocean ridge heterogeneities etc...). The other approach is to work with the theory of dynamic systems, because mantle convection is chaotic. The use of Lyapunov exponent characterize the local and global properties of stretching of heterogeneities. Mixing happens when an heterogeneity is stretched at a scale at which chemical diffusion operates. This scale for the Earth is about 1cm I would say (some papers talk about it). Lyapunov exponent are a form of stretching rate when mixing is chaotic. Configurational entropy and Lyapunov exponent can be related with Ergodicity theory. "

In Tackley and Xie [1314] (2002) we read:

> " Two-dimensional simulations of simple mantle convection (Christensen [262], 1989; Kellogg and Turcotte [731], 1990; Schmalzl and Hansen [1195]) suggest that for whole-mantle convection the mantle should be homogenized in a time-scale of less than 1 Gyr, although unmixed islands may remain. Non-Newtonian rheology, which is thought to be important in the upper mantle [714], may somewhat inhibit mixing (Ten, Podladchikov, Yuen, Larsen, and Malevsky [1329], 1998). The rate at which blobs of anomalous (e.g. primitive) material are stretched and assimilated into the flow depends on their relative viscosity; very viscous blobs can survive intact for many mantle overturns (Manga [881], 1996).
> In three-dimensional geometry with only poloidal flow, mixing may be significantly less efficient than in two dimensions (Schmalzl, Houseman, and Hansen [1197], 1995; Schmalzl, Houseman, and Hansen [1196], 1996) but, if the toroidal flow associated with plate motions is included (Gable, O'connell, and Travis [459], 1991), mixing can instead be more efficient (Ferrachat and Ricard [418] (1998)).
> High viscosity in the deep mantle is not sufficient to maintain different reservoirs over geological time-scales (Ferrachat and Ricard [417], 2001; van Keken & Ballentine 1998, 1999), in contrast to predictions from earlier calculations at lower convective vigour (Gurnis and Davies [545], 1986). Part of the reason for this apparent discrepancy is that the latter study used a kinematically driven flow rather than buoyancy-driven flow. Since a viscosity jump does not affect densities, thermal buoyancy-driven flow has no problem crossing it, so substantial mass exchange occurs between upper and lower mantles. Buoyancy is thus necessary to maintain separate reservoirs. "

In Gottschaldt, Walzer, Hendel, Stegman, Baumgardner, and Mühlhaus [504] (2006) we find

" Heterogeneities in a convecting fluid are deformed by stirring and finally erased by diffusive mixing. Chemical diffusion in mantle rock acts on the scale of centimetres over the lifetime of the Earth, but our models resolve km-scales. Therefore this paper deals only with convective stirring. Some nice studies about mantle stirring and mixing in 2-d have been done (e.g. [262, 545, 734, 731, 914, 1017, 1016, 1327, 1328, 1329]). Unfortunately, results from studies in 2-d can be extrapolated to 3-d only in a limited manner. Tracers in 3-d poloidal stationary convection move on 2-d toruslike surfaces. Stirring is constrained to these surfaces. Cross-cell stirring becomes possible in time-dependent flows, but may not be very efficient [1196]. Large-scale stirring is enhanced by a toroidal component, but convectively isolated islands of laminar stirring may remain [418]. Convection in the Earth is time-dependent and the surface planform shows a strong toroidal component today. Since there is currently no general understanding about the stirring behaviour of the mantle, a case-by-case study of different models is necessary. "

Kellogg and Stewart [734] (1991) looked into chaotic mixing and state:

" The extent of mixing is quantitatively measured by calculating the deformation under-gone by passive, elliptical strain markers. Deformation of a marker is characterized by its elongation a/a_0, where a is the length of the semimajor axis and a_0, is the initial length. For infinitesimal markers, deformation is controlled by the local flow field and is obtained by integrating the strain rate tensor along the particle's trajectory [731]. The aspect ratio ϵ and orientation θ of the ellipse vary with the nondimensional time according to

$$\frac{d\epsilon}{dt} = 2\epsilon(\bar{\alpha}\cos 2\theta + \bar{\beta}\sin 2\theta) \tag{16.83}$$

and

$$\frac{d\theta}{dt} = \left(\frac{\epsilon^2 + 1}{\epsilon^2 - 1}\right)(-\bar{\alpha}\sin 2\theta + \bar{\beta}\cos 2\theta) + \bar{\omega}. \tag{16.84}$$

The components of the nondimensional strain tensor $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\omega}$ are defined by

$$\bar{\alpha} = \frac{\partial u}{\partial x} \tag{16.85}$$

$$\bar{\beta} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \tag{16.86}$$

$$\bar{\omega} = \frac{1}{2}\left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right) \tag{16.87}$$

These equations define the evolution of an infinites'imal strain marker in an arbitrary, two-dimensional flow. Given the path of a particle in any flow, we can evaluatei $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\omega}$ at each position of the particle, and integrate (16.83) and (16.84) along the particle path to obtain the deformation history of a strain marker. These equations are derived in Kellogg [732] (1993). "

In Farnetani and Samuel [410] we find this important distinction:

" Stirring involves the repeated action of stretching and folding of a fluid element, while mixing also involves diffusion at small scales. The term stirring is more appropriate for the mantle, since rocks chemical diffusivity is only $10^{-15} - 10^{-20}$. "

In Samuel, Aleksandrov, and Deo [1176] (2011) we find:

" We measure the convective stirring efficiency using two Lagrangian methods: the first determines the mixing time associated with different wavelengths of heterogeneity following the approach of Ferrachat and Ricard [417] (2001).
The second determines the value of the maximum Finite Time Lyapunov Exponents (FTLE), λ^+ as described by Farnetani and Samuel [410] (2003), and measures the rate at which heterogeneities will be stretched by mantle motions. These methods were applied after each experiment had reached statistical steady state. "

Coltice [290] (2005) looked at convective mixing in degassing the Earth's mantle and describes his methodology as follows:

" To follow stretching and mimic the chemical evolution of the system, passive particles are tracked into the flow. Their position x into the convection box evolves according to

$$\frac{\partial x}{\partial t} = u(x, t),$$

u being the local velocity. To solve this equation, a time integration method and an interpolation of the velocity between grid nodes are needed. For sake of consistency, the velocity of a particle lying between grid nodes is naturally interpolated with the bilinear shape functions of the finite element formulation in ConMan. A fifth-order Runge–Kutta scheme is used for the time integration. In the simulations, I have checked that the number of particles is always statistically sufficient (increasing this number does not change the results).

Two simple models can describe the two end- members of mixing dynamics: regular (or laminar) and chaotic (or turbulent) mixing [1016]. Regular mixing occurs in 2D steady-state flows where the trajectories of particles are identical to streamlines. In this regime, an initial distance l_0 between two particles is stretched into a new distance l that increases linearly with time and proportionally to the strain rate. The stretching of a distance is parameterized by

$$\epsilon = l/l_0$$

which depends linearly on the time as

$$\epsilon(t) = 1 + \dot{\varepsilon}t \tag{16.88}$$

where $\dot{\varepsilon}$ is the Lagrangian strain rate. The Lagrangian strain rate is the strain rate experienced locally by the line as it is moving into the flow. In the calculations, the Lagrangian strain rate is always lower than the Eulerian strain rate which is the second invariant of the local strain rate tensor.

Chaotic mixing occurs when the dimensionality of the system is larger and where particle trajectories get more complex. In 2D, time-dependent systems show this behavior and two particles can follow completely different path-lines though they were infinitely close initially. In 3D, even steady flows can produce chaotic Lagrangian mixing, enhanced by the presence of a toroidal component [418]. Stretching in such a flow evolves exponentially with time

$$\epsilon(t) = \exp(\dot{\varepsilon}t)$$

The simplicity of these models is that they only depend on one parameter, the Lagrangian strain rate $\dot{\varepsilon}$. This parameter is chosen to be the quantitative measurement of average mixing in the numerical experiments.
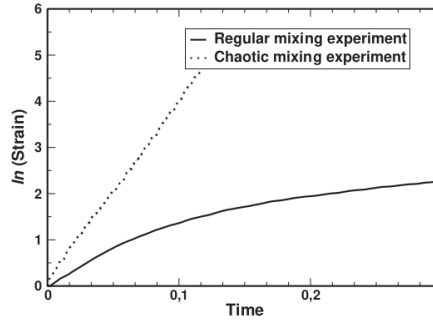
Strain as a function of time in two convection simulations illustrating regular mixing (internal heating, Ra i = 106) and chaotic mixing (basal heating, Ra b = 105).

As illustrated by the figure above, the numerical experiments display strain evolutions that can be well explained by one of the two end-members of mixing dynamics

To compute the local stretching ϵ around a particle, I follow the distance between two initially very close particles. The larger the distance, the larger the error on the local stretching. Hence, this distance has to be renormalized along the direction of stretching at every time step [418]. At a time t, the estimate of the Lagrangian strain rate is

$$\dot{\varepsilon} = \frac{\log(\epsilon)}{t}$$

The bulk Lagrangian strain rate is the average over all the couples of particles (20,000 were sufficient). In regular mixing flows, the computed ė would be much lower than the theoretical $\dot{\varepsilon}$ of Eq. (16.88). ”

Coltice and Schmalzl [292] (2006) show that mixing in 3-D time-dependent convection is as efficient as in 2-D, and only depends on convective vigor. compute mixing times of thermal and chemical heterogeneities in the early Earth of 10 and 100 Myrs respectively. They write:

“ To compute the Lagrangian strain rates, we follow the distance d(t) between particle twins advected in the flow using a 4th-order Runge-Kutta scheme. The Lagrangian strain rate, i.e. finite-time Lyapunov exponent, is computed assuming that for each couple of tracers, stretching $d(t)/d(0)$ occurs as

$$\frac{d(t)}{d(0)} = \exp(\dot{\varepsilon}t)$$

Averaging $\dot{\varepsilon}$ over the tracer couples gives the bulk Lagrangian strain rate characterizing the stretching ability of the flow [290]. If the computed bulk Lagrangian strain rate tends toward 0, mixing is regular, otherwise it is chaotic. To start the experiment, the distribution of the couples of tracers is homogeneous throughout the box. In the 3-D calculations, up to 10,000,000 tracers are used for an accurate description of the Lagrangian strain-rate field in 3-D. ”

Based on these, we need to look at the Lyapunov exponent (and the Finite Time Lyapunov Exponent, i.e. FTLE).

16.23.1   The Lyapunov exponent

Simply put, the Lyapunov time is the characteristic timescale on which a dynamical system is chaotic. It is defined as the inverse of a system's largest Lyapunov exponent.

The Lyapunov time mirrors the limits of the predictability of the system. By convention, it is defined as the time for the distance between nearby trajectories of the system to increase by a factor of e. However, measures in terms of 2-foldings and 10-foldings are sometimes found, since they correspond to the loss of one bit of information or one digit of precision respectively.

The Lyapunov exponent or Lyapunov characteristic exponent of a dynamical system is a quantity that characterizes the rate of separation of infinitesimally close trajectories. Quantitatively, two trajectories in phase space with initial separation $\delta\mathbf{Z}_0$ diverge (provided that the divergence can be treated within the linearized approximation) at a rate given by

$$|\delta\mathbf{Z}(t)| \approx e^{\lambda t}|\delta\mathbf{Z}_0|$$

where λ is the Lyapunov exponent.

Measuring the Lyapunov exponent or time (or related quantities) is relevant in the context of mantle stirring. On the one hand it is argued that the mantle is convecting and very efficient at mixing resulting in a somewhat homogenous composition. On the other hand, there is are modeling studies that suggest that whole-mantle convection can preserve heterogeneity in the presence of well-mixed mantle.

Mixing takes place by the repeated stretching and folding of interfaces. A measure of the mixing efficiency is the time evolution of the area of the mixing surface. Maximum efficiency of mixing is reached with turbulent mixing behavior where One can formally show whether mixing is laminar or turbulent by evaluating the Luyaponov exponents σ . These are of the form:

$$\sigma = \lim_{t \to \infty} \lim_{X \to 0} \left[\frac{1}{t} \ln\left(\frac{X(t)}{X(t=0)}\right)\right]$$

where $X(t)$ is the length of this segment at time t. Non-zero Luyaponov exponents indicate that stretching is exponential and the larger the exponent, the more efficient mixing is. However, the limits in the above equation are difficult to evaluate and the interpretation of the 'finite-time' Luyaponov exponent, where both limits are truncated, is difficult to formalize.

In van Keken and Zhong [1405] (1999) the authors use a steady state velocity pattern obtained for a model of present-day mantle convection. The velocity model is based on the solution of the Stokes equations in a 3D spherical model with variable rheology. To study mixing, they release particles in the velocity model and follow these by numerical integration.
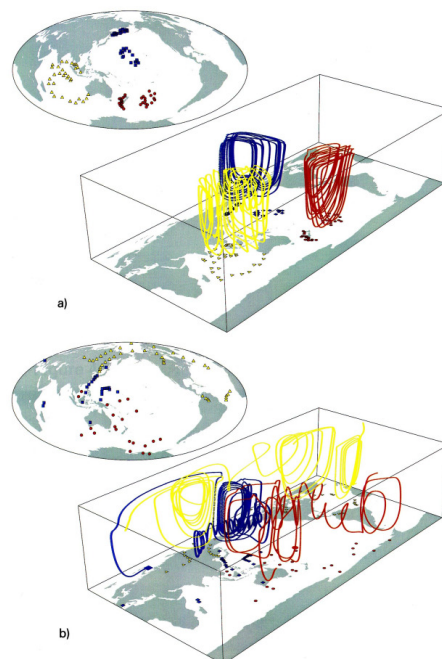
a) The three particles in this plot were selected for their relatively regular pattern. b) Three other particles that traverse a large portion of the model. These particles feel the strong toroidal motion and their paths form corkscrew-like patterns. They indicate that certain parts of the model can exhibit strong mixing. Taken from van Keken and Zhong [1405] (1999).

Rather than calculating the exponents explicitly, the authors use an approximation to the finite-time, finite-length Lyaponov exponent by evaluating the distance between two points that are closely spaced at time $t = 0$. For this they compute the advection of a large number of 10 km long line segments that were originally at 1500 km depth. The length of these segments is approximated by the distance between the endpoints and the results are summarized in the following figure:
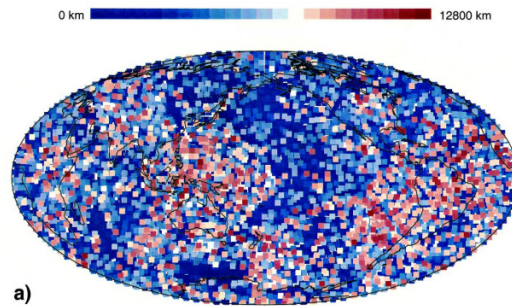
Length of the line segment after 4 billion years. Approximately 14,000 line segments were released with regular spacing at 1500 km depth. The length of the segment is indicated by the colored symbols that are plotted at the initial position. The results indicate that there is a strong diversity in mixing behavior. In some regions (north Pacific, parts under the Indian/Australian plate) stretching is very limited, indicating laminar and consequently inefficient mixing. Regions that are under strong toroidal surface motion (western Pacific, Nazca and South America) show very efficient stretching of up to the maximum length of the diameter of the Earth. Taken from van Keken and Zhong [1405] (1999).

In Bello, Coltice, Rolf, and Tackley [75] (2014) the authors estimate for the first time the limit of predictability of Earth's mantle convection. Following the twin experiment method, we compute the Lyapunov time (i.e., e-folding time) for state of the art 3-D spherical convection models, varying rheology, and Rayleigh number.

Reconstruction of mantle convection and surface tectonics with (ensemble) Kalman filter: Bocher, Coltice, Fournier, and Tackley [106] (2016), Bocher, Fournier, and Coltice [107] (2018).

Investigating the initial condition of mantle models using data assimilation. PhD thesis. Price [1081] (2016).

📦 Relevant Literature

Pierrehumbert [1060] (1991), Colli, Bunge, and Schuberth [289] (2015), Cagney, Newsome, Lithgow-Bertelloni, Cotel, Hart, and Whitehead [216] (2015), Thomas, Samuel, Farnetani, Aubert, and Chauvel [1352] (2024).

16.23.2 Following patches of passively advected markers

In Gurnis [543] (1986) the stirring of a small, passive heterogeneity in an unsteady circulation is investigated by following the boundary of the heterogeneity; the circulation mimics features of plate-scale flow. After initial release, the heterogeneity is stirred and subsequently consists of at least one large "blob" connected to long, but thin tendrils.
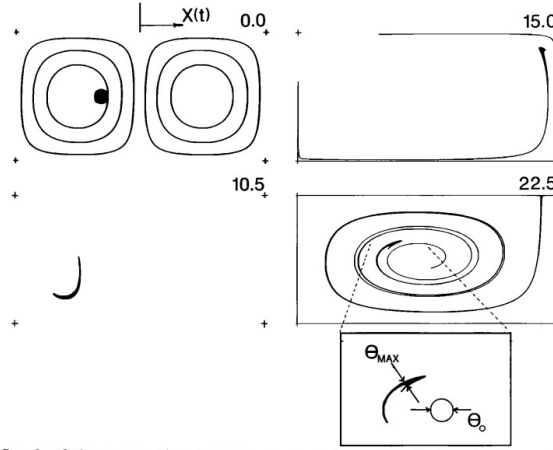
Fig. 2. Example of the stirring of an initially circular heterogeneity, black patch, in a periodic flow. The streamlines at time 0 are shown; the margin between the two cells oscillates and the flow is shown at many more steps in Gurnis and Davies (1986). Each frame is labelled with transit times. The inset highlights a large and persistent heterogeneity connected to very long, but thin tendrils. The maximum width is on the order of the initial size of the heterogeneity. Borders on the 2x1 boxes have not been drawn: any line on the edge of a box is a tendril.

16.23.3   Hyperbolic persistence time method

This method is presented in Farnetani and Samuel [410] (2003). Looks complex, and too long to copy paste here.

16.23.4   configurational 'Shannon' entropy

Goltz and Böse [501] (2002), Camesasca, Kaufman, and Manas-Zloczower [220] (2006), Naliboff and Kellogg [981] (2007), Perugini, De Campos, Petrelli, Morgavi, Vetere, and Dingwell [1054] (2015), Bianchi and Pedretti [94] (2018), Wiel, Thieulot, and Hinsbergen [1455] (2024).

16.24 Phase transitions

The topic of phase transitions and their implementation is computational geodynamics is a *very* vast topic. It requires input from thermodynamics, geochemistry and petrology, and also requires dedicated algorithms which are quite complex.

Let us start with simple examples from the literature:

- Zlotnik *et al.* (2007) [1547]. The equations that is used in this work are the standard incompressible Stokes equations by the authors chose to represent the density as a function of of temperature and pressure by the following expression:

$$\rho(T, p) = \rho_0[1 - \alpha(T - T_0)][1 + \beta(p - p_0)]$$

where α and β are, respectively, the thermal expansion and compressibility coefficients, and T_0 and p_0 are reference values at surface.

The authors then proceed to divide the phase diagram into three regions corresponding to three minerals: olivine, spinel-structured olivine, and perovskite:
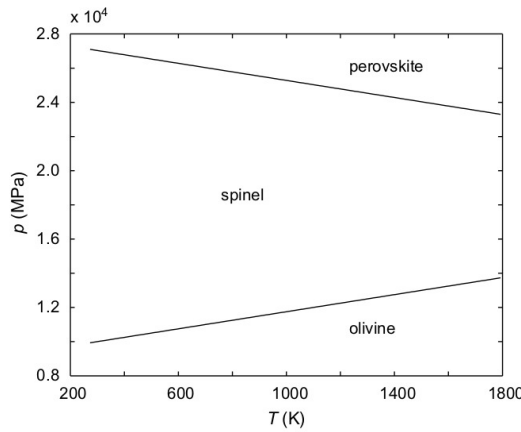
Phase diagram indicating stable mineral phases in the temperature-pressure plane. The phase diagram is divided into three regions corresponding to three distinct minerals: olivine, spinel and perovskite. Taken from [1547].

They state that two major mineralogical phase transitions occur, one at 410 km depth and other at 660 km depth (other deeper transitions run outside the domain under study because their domain is 1000km deep). The density increases discontinuously across these phase transitions. In order to take into account the effect of these discontinuities, the density ρ_0 above is taken as a reference density plus an increment $\Delta\rho$:

$$\rho_0 = \rho_{olivine} + \Delta\rho$$

where

$$\Delta\rho = \begin{cases} 0 & \text{if } (T, p)\text{is in the olivine region} \\ \Delta\rho_{es} & \text{if } (T, p)\text{is in the spinel region} \\ \Delta\rho_{per} & \text{if } (T, p)\text{is in the perovskite region} \end{cases}$$

The authors unfortunately fail to report how the phase transitions affect the viscosity.

The obvious problem with this otherwise simple approach is that density varies in the domain but is not accompanied by a volume change so that it violates mass conservation.

- the following phase diagram is taken from Peltier *et al.* (1997) [1049].
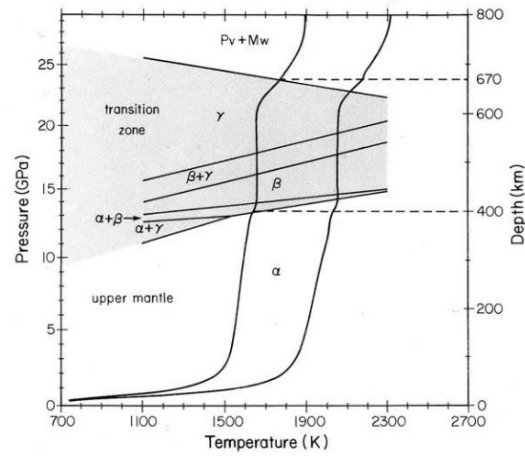
Phase boundary for the $\alpha \to \beta \to \gamma$ transitions of Olivine
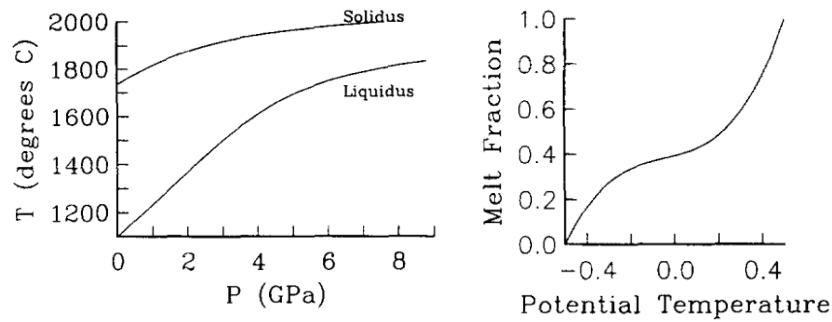
Left: Solidus and liquidus temperatures; Right: melt fraction as a function temperature for dry peridotite. Taken from King & Anderson (1995) [744], Both after McKenzie & Bickle [903].
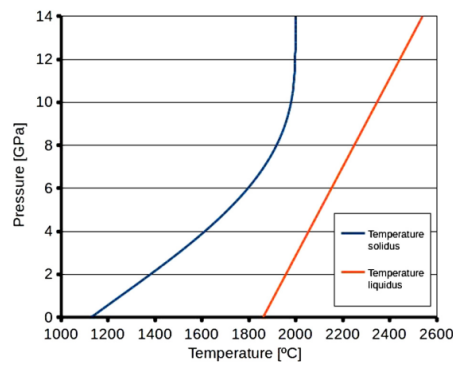
Left: Solidus and liquidus curves. Taken from Lavecchia *et al.* (2017) [798].

📚 Relevant Literature: [1220]

16.25 Implementation of an elasto-viscous rheology

A viscoelastic material can behave both elastically and viscously. Its response to an applied stress is dependent on the material properties and can be found using the Maxwell time (t_M) of said material. This material constant is defined as the ratio of the material viscosity and shear modulus

$$t_M = \frac{\eta}{\mu}$$

where η is the viscosity and μ the shear modulus.

The total deviatoric strainrate tensor can be decomposed into an elastic component and a viscous component:

$$\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}) = \dot{\boldsymbol{\varepsilon}}^d_e(\vec{\mathbf{v}}) + \dot{\boldsymbol{\varepsilon}}^d_v(\vec{\mathbf{v}}) = \frac{\tilde{\dot{\boldsymbol{\tau}}}}{2\mu} + \frac{\boldsymbol{\tau}}{2\eta}$$

[From wikipedia] In continuum mechanics, objective stress rates are time derivatives of stress that do not depend on the frame of reference. Many constitutive equations are designed in the form of a relation between a stress-rate and a strain-rate (or the rate of deformation tensor). i The mechanical response of a material should not depend on the frame of reference. In other words, material constitutive equations should be frame indifferent (objective). If the stress and strain measures are material quantities then objectivity is automatically satisfied. However, if the quantities are spatial, then the objectivity of the stress-rate is not guaranteed even if the strain-rate is objective.

There are numerous objective stress rates in continuum mechanics - all of which can be shown to be special forms of Lie derivatives. Some of the widely used objective stress rates are [619]: a) the Truesdell rate of the Cauchy stress tensor, b) the Green-Naghdi rate of the Cauchy stress, and c) the Jaumann rate of the Cauchy stress.

The Jaumann rate of the Cauchy stress is a further specialization of the Lie derivative (Truesdell rate). This rate has the form

$$\tilde{\dot{\boldsymbol{\tau}}}^{t+\delta t} = \frac{D\boldsymbol{\tau}}{Dt} - (\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)) = \frac{\boldsymbol{\tau}^{t+\delta t} - \boldsymbol{\tau}^t}{\delta t} - (\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t))$$

where D/Dt is the material derivative and $\dot{\boldsymbol{\omega}}$ is the rotation rate -also called spin tensor- which is anti-symmetric and has zero trace - see Section 3.4.3:

$$\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathbf{v}} - (\vec{\nabla}\vec{\mathbf{v}})^T \right)$$

In the case of a Lagrangian description, we have [1428]

$$\tilde{\dot{\boldsymbol{\tau}}}^{t+\delta t} = \frac{\boldsymbol{\tau}^{t+\delta t} - \boldsymbol{\tau}^t}{\delta t} - (\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t))$$

so that

$$\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) = \frac{\tilde{\dot{\boldsymbol{\tau}}}^{t+\delta t}}{2\mu} + \frac{\boldsymbol{\tau}^{t+\delta t}}{2\eta} = \frac{1}{2\mu}\left[\frac{\boldsymbol{\tau}^{t+\delta t} - \boldsymbol{\tau}^t}{\delta t} - (\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)) \right] + \frac{\boldsymbol{\tau}^{t+\delta t}}{2\eta}$$

Let us multiply this by $2\mu\delta t$ and transform the equations until a satisfying formulation is found:

$$2\mu\delta t\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) = \boldsymbol{\tau}^{t+\delta t} - \boldsymbol{\tau}^t - \delta t(\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)) + \frac{\mu\delta t}{\eta}\boldsymbol{\tau}^{t+\delta t}$$

$$2\mu\delta t\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) = \left(1 + \frac{\mu\delta t}{\eta} \right)\boldsymbol{\tau}^{t+\delta t} - \boldsymbol{\tau}^t - \delta t(\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t))$$

$$\left(1 + \frac{\mu\delta t}{\eta}\right)\boldsymbol{\tau}^{t+\delta t} = 2\mu\delta t\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + \boldsymbol{\tau}^t + \delta t(\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)\cdot\boldsymbol{\tau}^t - \boldsymbol{\tau}^t\cdot\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t))$$

$$\boldsymbol{\tau}^{t+\delta t} = 2\frac{\mu\delta t}{\left(1 + \frac{\mu\delta t}{\eta}\right)}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + \frac{1}{\left(1 + \frac{\mu\delta t}{\eta}\right)}\boldsymbol{\tau}^t + \frac{\delta t}{\left(1 + \frac{\mu\delta t}{\eta}\right)}(\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)\cdot\boldsymbol{\tau}^t - \boldsymbol{\tau}^t\cdot\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t))$$

We define:

$$\boxed{\eta_{eff} = \frac{\mu\delta t}{\left(1 + \frac{\mu\delta t}{\eta}\right)} = \frac{\eta\delta t}{\delta t + \eta/\mu} = \frac{\eta}{1 + t_M/\delta t}} \tag{16.89}$$

$$\boxed{Z = \frac{\eta_{eff}}{\mu\delta t} = \frac{\eta}{\mu\delta t + \eta}}$$

$$\boxed{\boldsymbol{J}^t = \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)\cdot\boldsymbol{\tau}^t - \boldsymbol{\tau}^t\cdot\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t)}$$

so that we can write

$$\boldsymbol{\tau}^{t+\delta t} = 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + Z\boldsymbol{\tau}^t + Z\delta t\boldsymbol{J}^t$$

or,

$$\boldsymbol{\tau}^{t+\delta t} = 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + \underline{\boldsymbol{\tau}}^t \qquad \text{with} \qquad \underline{\boldsymbol{\tau}}^t = Z\boldsymbol{\tau}^t + Z\delta t\boldsymbol{J}^t$$

The total stress tensor is then

$$\boxed{\boldsymbol{\sigma}^{t+\delta t} = -p^{t+\delta t}\mathbf{1} + \boldsymbol{\tau}^{t+\delta t} = -p^{t+\delta t}\mathbf{1} + 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + Z\boldsymbol{\tau}^t + Z\delta t\boldsymbol{J}^t} \tag{16.90}$$

Remark. *When $\mu \to \infty$ we have $\eta_{eff} \to \eta$ and $Z \to 0$ and we recover the Stokes equation for a purely viscous fluid.*

Strong form

Let us now turn to the momentum conservation equation:

$$\vec{\nabla}\cdot\boldsymbol{\sigma}^{t+\delta t} + \rho^{t+\delta t}\vec{g} = \vec{0}$$
$$\Rightarrow \quad \vec{\nabla}\cdot(-p^{t+\delta t}\mathbf{1} + \boldsymbol{\tau}^{t+\delta t}) + \rho^{t+\delta t}\vec{g} = \vec{0}$$
$$\Rightarrow \quad -\vec{\nabla}p^{t+\delta t} + \vec{\nabla}\cdot\boldsymbol{\tau}^{t+\delta t} + \rho^{t+\delta t}\vec{g} = \vec{0}$$
$$\Rightarrow \quad -\vec{\nabla}p^{t+\delta t} + \vec{\nabla}\cdot\left(2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + \underline{\boldsymbol{\tau}}^t\right) + \rho^{t+\delta t}\vec{g} = \vec{0}$$

and finally

$$\boxed{-\vec{\nabla}p^{t+\delta t} + \vec{\nabla}\cdot 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) = -\rho^{t+\delta t}\vec{g} - \vec{\nabla}\cdot\underline{\boldsymbol{\tau}}^t}$$

Weak form

The mass conservation equation is still $\vec{\nabla}\cdot\vec{\mathbf{v}} = 0$ so we need not look into it since its weak form is in Section 10.8.

For the N_i^v's we can write:

$$\int_{\Omega_e} N_i^v \vec{\nabla}\cdot\boldsymbol{\sigma}^{t+\delta t}d\Omega + \int_{\Omega_e} N_i^v \rho\vec{g}\,d\Omega = \vec{0} \tag{16.91}$$

We can integrate by parts and drop the surface term[33] REVISIT and use Eq. (16.90):

$$\int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \boldsymbol{\sigma}^{t+\delta t} \, d\Omega \;=\; \int_{\Omega_e} N_i^\gamma \rho \vec{g} \, d\Omega$$

$$\int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \left[-p^{t+\delta t}\mathbf{1} + 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) + \underline{\boldsymbol{\tau}}^t \right] \, d\Omega \;=\; \int_{\Omega_e} N_i^\gamma \rho \vec{g} \, d\Omega \tag{16.92}$$

$$\int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \left[-p^{t+\delta t}\mathbf{1} + 2\eta_{eff}\dot{\boldsymbol{\varepsilon}}^d(\vec{\mathbf{v}}^{t+\delta t}) \right] \, d\Omega \;=\; \int_{\Omega_e} N_i^\gamma \rho \vec{g} \, d\Omega - \int_{\Omega_e} \vec{\nabla} N_i^\gamma \cdot \underline{\boldsymbol{\tau}}^t \, d\Omega \tag{16.93}$$

We see that the left hand term is virtually identical to the one in Section 10.8, although the viscosity has been replaced with the effective viscosity of Eq. (16.89). The headache will come from the right hand side term $\underline{\boldsymbol{\tau}}^t$, as we will see.

In two dimensions - Cartesian coordinates . The rotation rate tensor is given by:

$$\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathbf{v}} - (\vec{\nabla}\vec{\mathbf{v}})^T \right) = \begin{pmatrix} 0 & \dot{\omega}_{xy} \\ -\dot{\omega}_{xy} & 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 0 & \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} & 0 \end{pmatrix}$$

so that the tensor $\boldsymbol{J}$ can be computed explicitly:

$$
\begin{aligned}
\boldsymbol{J}^t &= \dot{\boldsymbol{\omega}}(\vec{v}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{v}^t) \\
&= \begin{pmatrix} 0 & \dot{\omega}_{xy}(\vec{v}^t) \\ -\dot{\omega}_{xy}(\vec{v}^t) & 0 \end{pmatrix} \cdot \begin{pmatrix} \tau_{xx}^t & \tau_{xy}^t \\ \tau_{xy}^t & \tau_{yy}^t \end{pmatrix} - \begin{pmatrix} \tau_{xx}^t & \tau_{xy}^t \\ \tau_{xy}^t & \tau_{yy}^t \end{pmatrix} \cdot \begin{pmatrix} 0 & \dot{\omega}_{xy}(\vec{v}^t) \\ -\dot{\omega}_{xy}(\vec{v}^t) & 0 \end{pmatrix} \\
&= \dot{\omega}_{xy}(\vec{v}^t)\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \tau_{xx}^t & \tau_{xy}^t \\ \tau_{xy}^t & \tau_{yy}^t \end{pmatrix} - \dot{\omega}_{xy}(\vec{v}^t)\begin{pmatrix} \tau_{xx}^t & \tau_{xy}^t \\ \tau_{xy}^t & \tau_{yy}^t \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\
&= \dot{\omega}_{xy}(\vec{v}^t)\begin{pmatrix} \tau_{xy}^t & \tau_{yy}^t \\ -\tau_{xx}^t & -\tau_{xy}^t \end{pmatrix} - \dot{\omega}_{xy}(\vec{v}^t)\begin{pmatrix} -\tau_{xy}^t & \tau_{xx}^t \\ -\tau_{yy}^t & \tau_{xy}^t \end{pmatrix} \\
&= \dot{\omega}_{xy}(\vec{v}^t)\begin{pmatrix} 2\tau_{xy}^t & \tau_{yy}^t - \tau_{xx}^t \\ \tau_{yy}^t - \tau_{xx}^t & -2\tau_{xy}^t \end{pmatrix} \tag{16.94}
\end{aligned}
$$

so that the tensor equation $\underline{\boldsymbol{\tau}} = Z\boldsymbol{\tau} + Z\delta t \boldsymbol{J}$ can be reformulated as follows in a vector form:

$$\begin{pmatrix} \underline{\tau}_{xx}^t \\ \underline{\tau}_{yy}^t \\ \underline{\tau}_{xy}^t \end{pmatrix} = Z\begin{pmatrix} \tau_{xx}^t \\ \tau_{yy}^t \\ \tau_{xy}^t \end{pmatrix} + Z\delta t\begin{pmatrix} J_{xx}^t \\ J_{yy}^t \\ J_{xy}^t \end{pmatrix} = Z\begin{pmatrix} \tau_{xx}^t \\ \tau_{yy}^t \\ \tau_{xy}^t \end{pmatrix} + Z\delta t\dot{\omega}_{xy}^t\begin{pmatrix} 2\tau_{xy}^t \\ -2\tau_{xy}^t \\ \tau_{yy}^t - \tau_{xx}^t \end{pmatrix} \tag{16.95}$$

or,

$$
\begin{aligned}
\begin{pmatrix} \sigma_{xx}^{t+\delta t} \\ \sigma_{yy}^{t+\delta t} \\ \sigma_{xy}^{t+\delta t} \end{pmatrix} &= \begin{pmatrix} -p^{t+\delta t} \\ -p^{t+\delta t} \\ 0 \end{pmatrix} + 2\eta_{eff}\begin{pmatrix} \dot{\varepsilon}_{xx}(\vec{v}^{t+\delta t}) \\ \dot{\varepsilon}_{yy}(\vec{v}^{t+\delta t}) \\ \dot{\varepsilon}_{xy}(\vec{v}^{t+\delta t}) \end{pmatrix} + \begin{pmatrix} \underline{\tau}_{xx}^t \\ \underline{\tau}_{yy}^t \\ \underline{\tau}_{xy}^t \end{pmatrix} \\
&= \begin{pmatrix} -p^{t+\delta t} \\ -p^{t+\delta t} \\ 0 \end{pmatrix} + 2\eta_{eff}\begin{pmatrix} \dot{\varepsilon}_{xx}(\vec{v}^{t+\delta t}) \\ \dot{\varepsilon}_{yy}(\vec{v}^{t+\delta t}) \\ \dot{\varepsilon}_{xy}(\vec{v}^{t+\delta t}) \end{pmatrix} + Z\begin{pmatrix} \tau_{xx}^t \\ \tau_{yy}^t \\ \tau_{xy}^t \end{pmatrix} + Z\delta t\dot{\omega}_{xy}\begin{pmatrix} 2\tau_{xy}^t \\ -2\tau_{xy}^t \\ \tau_{yy}^t - \tau_{xx}^t \end{pmatrix} \tag{16.96}
\end{aligned}
$$

...

$$\int_{\Omega_e} \boldsymbol{B}^T \cdot \begin{pmatrix} \sigma_{xx}^{t+\delta t} \\ \sigma_{yy}^{t+\delta t} \\ \sigma_{xy}^{t+\delta t} \end{pmatrix} d\Omega = \int_{\Omega_e} \vec{N}_b d\Omega \tag{16.97}$$

---

[33] We will come back to this at a later stage

$$
\int_{\Omega_e} \boldsymbol{B}^T \cdot \left[\begin{pmatrix} -p^{t+\delta t} \\ -p^{t+\delta t} \\ 0 \end{pmatrix} + 2\eta_{eff} \begin{pmatrix} \dot{\varepsilon}^d_{xx}(\vec{\mathbf{v}}^{t+\delta t}) \\ \dot{\varepsilon}^d_{yy}(\vec{\mathbf{v}}^{t+\delta t}) \\ \dot{\varepsilon}^d_{xy}(\vec{\mathbf{v}}^{t+\delta t}) \end{pmatrix} \right] d\Omega = \int_{\Omega_e} \vec{N}_b d\Omega - \int_{\Omega_e} \boldsymbol{B}^T \cdot \left[Z \begin{pmatrix} \tau^t_{xx} \\ \tau^t_{yy} \\ \tau^t_{xy} \end{pmatrix} + Z\delta t \dot{\omega}_{xy}(\vec{\mathbf{v}}^t) \begin{pmatrix} 2\tau^t_{xy} \\ -2\tau^t_{x} \\ \tau^t_{yy} - \end{pmatrix} \right.
$$

As seen in Section 10.8 the left hand side terms yield $\mathbb{K} \cdot \vec{V} + \mathbb{G} \cdot \vec{P}$. The buoyancy term in the rhs is also standard and yields $\vec{f}$. The discretised momentum equation then writes

$$
\mathbb{K} \cdot \vec{V} + \mathbb{G} \cdot \vec{P} = \vec{f} + \vec{f}_{el}
$$

and the last rhs term is

$$
\boxed{\vec{f}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \left[Z \begin{pmatrix} \tau^t_{xx} \\ \tau^t_{yy} \\ \tau^t_{xy} \end{pmatrix} + Z\delta t \dot{\omega}_{xy}(\vec{\mathbf{v}}^t) \begin{pmatrix} 2\tau^t_{xy} \\ -2\tau^t_{xy} \\ \tau^t_{yy} - \tau^t_{xx} \end{pmatrix} \right] d\Omega}
$$

with the matrix $\boldsymbol{B}$ being given by

$$
\boldsymbol{B} = \begin{pmatrix} \frac{\partial N^{\vee}_1}{\partial x} & 0 & 0 & \cdots & \frac{\partial N^{\vee}_{m_v}}{\partial x} & 0 & 0 \\ \\ 0 & \frac{\partial N^{\vee}_1}{\partial y} & 0 & \cdots & 0 & \frac{\partial N^{\vee}_{m_v}}{\partial y} & 0 \\ \\ \frac{\partial N^{\vee}_1}{\partial y} & \frac{\partial N^{\vee}_1}{\partial x} & 0 & \cdots & \frac{\partial N^{\vee}_{m_v}}{\partial x} & \frac{\partial N^{\vee}_{m_v}}{\partial x} & 0 \end{pmatrix}
$$

In three dimensions - Cartesian coordinates The spin tensor is given by

$$
\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathbf{v}} - (\vec{\nabla}\vec{\mathbf{v}})^T \right) = \begin{pmatrix} 0 & \dot{\omega}_{xy} & \dot{\omega}_{xz} \\ -\dot{\omega}_{xy} & 0 & \dot{\omega}_{yz} \\ -\dot{\omega}_{xz} & -\dot{\omega}_{yz} & 0 \end{pmatrix}
$$

so that

$$
\begin{aligned}
\boldsymbol{J}^t &= \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \cdot \boldsymbol{\tau}^t - \boldsymbol{\tau}^t \cdot \dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}^t) \\
&= \begin{pmatrix} 0 & \dot{\omega}_{xy} & \dot{\omega}_{xz} \\ -\dot{\omega}_{xy} & 0 & \dot{\omega}_{yz} \\ -\dot{\omega}_{xz} & -\dot{\omega}_{yz} & 0 \end{pmatrix} \cdot \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{pmatrix} - \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{pmatrix} \cdot \begin{pmatrix} 0 & \dot{\omega}_{xy} & \dot{\omega}_{xz} \\ -\dot{\omega}_{xy} & 0 & \dot{\omega}_{yz} \\ -\dot{\omega}_{xz} & -\dot{\omega}_{yz} & 0 \end{pmatrix} \\
&= \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (16.98)
\end{aligned}
$$

FINISH!!!

check appendix A of Loes' GR

16.26 Nodal interpolation inside an element

The $n+1$ Bernstein basis polynomials of degree n on the interval $[0,1]$ are defined as [34]

$$b_{m,n}(x) = \left(\begin{array}{c} n \\ m \end{array} \right) x^m (1-x)^{n-m} \qquad m = 0, 1, ...n$$

The first few Bernstein polynomials are

$$
\begin{aligned}
b_{0,0}(x) &= 1 &\text{(16.99)} \\
b_{0,1}(x) &= 1 - x & \\
b_{1,1}(x) &= x &\text{(16.100)} \\
b_{0,2}(x) &= (1-x)^2 & \\
b_{1,2}(x) &= 2x(1-x) & \\
b_{2,2}(x) &= x^2 &\text{(16.101)}
\end{aligned}
$$
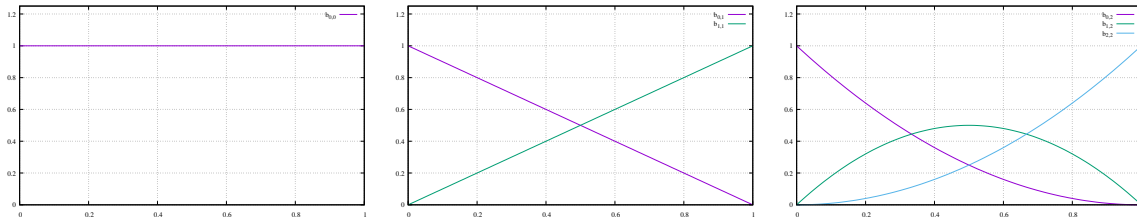
We see that the zero-th and first order polynomials are the same as the linear basis functions defined in Section 8.2. However the second order polynomials (and higher) differ from the second-order basis functions.

Also, the Bernstein polynomials have a lot of properties, but one that is of importance to us is the following: $b_{m,n}(x) \geq 0 \quad \forall x \in [0,1]$, i.e. the polynomials are positive. This is however not true for basis functions for $n \geq 2$. Another important property shared with basis functions is that their sum over the interval is exactly 1, i.e. $\sum_m b_{m,n}(x) = 1$.

In order to facilitate the comparison between the 2nd-order basis functions and Bernstein polynomials, I will express the latter as a function of the reduced coordinate $r \in [-1,1] = 2(x - 1/2)$ (or $x = (r+1)/2$). We have then:

$$
\begin{aligned}
b_{0,2}(r) &= \frac{1}{4}(1-r)^2 \\
b_{1,2}(r) &= \frac{1}{2}(1-r^2) \\
b_{2,2}(r) &= \frac{1}{4}(1+r)^2 &\text{(16.102)}
\end{aligned}
$$

Both 2nd-order Bernstein polynomials and basis functions are plotted here under:

---
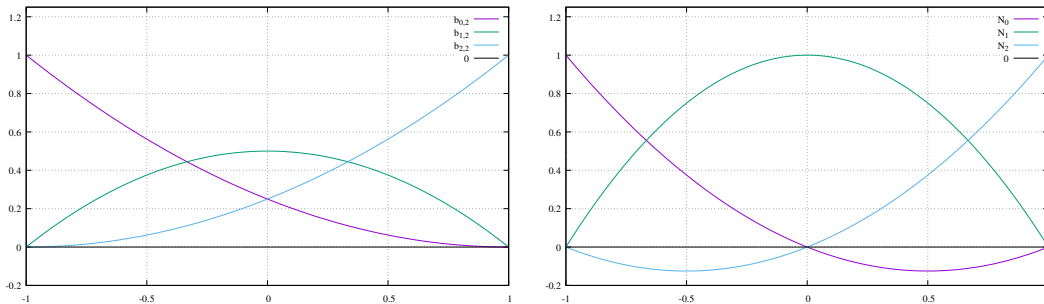
[34]https://en.wikipedia.org/wiki/Bernstein_polynomial

Having reached this point, the burning question is why should we care?

Example 1 In order to answer this question, let us carry out the following experiment: each node i in the element carries a field value f_i and for simplicity, we choose $f_0 = f(r = -1) = 1$, $f_1 = f(r = 0) = 0$, $f_2 = f(r = +1) = 0$. Then, we can compute the value of the field inside of the element as we usually do in the FE methodology:

$$f^h(r) = \sum_{i=0}^{2} \mathcal{N}_i(r) f_i = f_0 \mathcal{N}_0(r) = \mathcal{N}_0(r) = \frac{1}{2} r(r - 1)$$

This means that although the field f is always positive (or null) inside the element its representation with the basis functions is negative over half (!) of the element (see purple curve on the right panel above). If we now turn to the Bernstein polynomials:

$$f^h(r) = \sum_{i=0}^{2} b_{i,2} f_i = b_{0,2} = \frac{1}{4}(1 - r)^2$$

which is *always* positive over the interval $[-1, +1]$, and looking at the purple curve on the left panel above, we see that the value decreases monotonously when we go away from node 1, and reaches zero at the other end of the element.

Also:

$$\text{Shape function: } \int_{-1}^{+1} f^h(r) dr = \int_{-1}^{+1} \frac{1}{2} r(r - 1) dr = \frac{1}{3}$$

$$\text{Bernstein polynomial: } \int_{-1}^{+1} f^h(r) dr = \int_{-1}^{+1} \frac{1}{4}(1 - r)^2 dr = \frac{2}{3}$$

Analytical value for the integral can be obtained by splitting the integral as $\int_{-1}^{0} + \int_{0}^{+1}$. The left part can be represented by a line of equation $-r$ and the right part simply by 0, so that the integral is equal to $1/2$. We then see that the Shape function-based interpolation underestimates the integral while the Bernstein polynomial-based interpolation overestimates it.

Example 2 We now choose $f_0 = f(r = -1) = 1$, $f_1 = f(r = 0) = 1$, $f_2 = f(r = 1) = 0$. Then

$$f^h(r) = \sum_{i=0}^{2} f_i \mathcal{N}_i(r) = f_0 \mathcal{N}_0(r) + f_1 \mathcal{N}_1(r) = \mathcal{N}_0(r) + \mathcal{N}_1(r) = \frac{1}{2} r(r - 1) + 1 - r^2 = -\frac{1}{2} r^2 - \frac{1}{2} r + 1$$

Looking now at the Bernstein polynomials:

$$f^h(r) = \sum_{i=0}^{2} b_{i,2} f_i = b_{0,2} + b_{1,2} = \frac{1}{4}(1 - r)^2 + \frac{1}{2}(1 - r^2) = -\frac{1}{4} r^2 - \frac{1}{2} r + \frac{3}{4}$$

If we now plot both approximations:

795
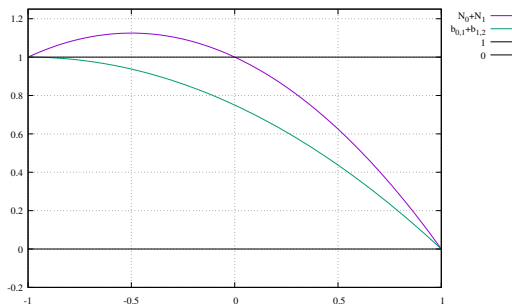
We see that in this case the Q_2 basis function-based approximation yields values above 1 over half of the element while the Bernstein polynomial-based approximation remains between 0 and 1 as expected.

Approximation of polynomials Let us now explore another aspect of such an interpolation based on the Bernstein polynomials and assume that $f(r) = C$. Then

$$f^h(r) = \sum_{i=0}^{2} f_i b_{i,2}(r) = C \sum_{i=0}^{2} b_{i,2}(r) = C \cdot 1 = C$$

Such interpolation can exactly represent a constant field. Let us assume that $f(r) = ar + b$. Then

$$
\begin{aligned}
f^h(r) &= \sum_{i=0}^{2} f(r_i) b_{i,2}(r) \\
&= \sum_{i=0}^{2} (ar_i + b) b_{i,2}(r) \\
&= a \sum_{i=0}^{2} r_i b_{i,2}(r) + b \sum_{i=1}^{3} b_{i,2}(r) \\
&= a(-b_{0,2}(r) + b_{2,2}(r)) + b \cdot 1 \\
&= ar + b
\end{aligned}
\tag{16.103}
$$

Such interpolation can exactly represent a linear field.

Let us assume that $f(r) = ar^2 + br + c$. Then

$$
\begin{aligned}
f^h(r) &= \sum_{i=0}^{2} f(r_i) b_{i,2}(r) \\
&= \sum_{i=0}^{2} (ar^2 + br + c) b_{i,2}(r) \\
&= \sum_{i=0}^{2} r_i^2 b_{i,2}(r) + b \sum_{i=0}^{2} r_i b_{i,2}(r) + c \sum_{i=0}^{2} b_{i,2}(r) \\
&= a \sum_{i=0}^{2} r_i^2 b_{i,2}(r) + br + c \\
&= a(b_{0,2}(r) + b_{2,2}(r)) + br + c \\
&= a\frac{1}{2}(1 + r^2) + br + c
\end{aligned}
\tag{16.104}
$$

which is not equal to $f(r)$.

On the other hand it is trivial to show that

$$
\begin{aligned}
f^h(r) &= \sum_{i=0}^{2} f(r_i)\mathcal{N}_i(r) \\
&= \sum_{i=0}^{2} (ar_i^2 + br_i + c)\mathcal{N}_i(r) \\
&= a\sum_{i=0}^{2} r_i^2 \mathcal{N}_i(r) + b\sum_{i=0}^{2} r_i \mathcal{N}_i(r) + c\sum_{i=0}^{2} \mathcal{N}_i(r) \\
&= a\sum_{i=0}^{2} r_i^2 \mathcal{N}_i(r) + b\sum_{i=0}^{2} r_i \mathcal{N}_i(r) + c \\
&= a(N_0(r) + N_2(r)) + b(-\mathcal{N}_0(r) + \mathcal{N}_2(r)) + c \\
&= ar^2 + br + c
\end{aligned}
\tag{16.105}
$$

To hammer the point once more: let $f(r) = r^2 + r + 1$. Then

$$
\begin{aligned}
f^h_{Q_1} &= f(-1)\frac{1}{2}(1 - r) + f(+1)\frac{1}{2}(1 + r) \\
&= \frac{1}{2}(1 - r) + 3\frac{1}{2}(1 + r) \\
&= 2 - r \tag{16.106} \\
f^h_{Q_2} &= r^2 + r + 1 \tag{16.107} \\
f^h_{B_2} &= \frac{1}{2}(1 + r^2) + r + 1 \tag{16.108}
\end{aligned}
$$

We see on the following figure that although Bernstein polynomials cannot represent $f(r)$ exactly they still do a better job than first order basis functions (Q_1 projection).
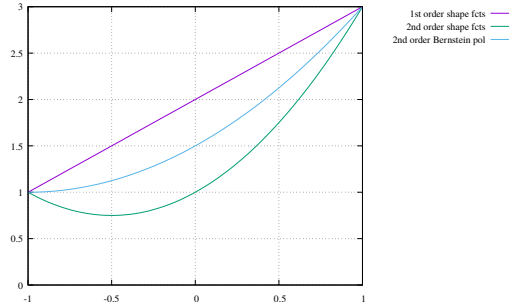
As a conclusion there is a trade-off: 2nd-order Bernstein polynomials *always* yield positive values when the field is positive (as opposed to 2nd-order basis functions) but they cannot represent exactly a 2nd-order polynomial field (while basis functions can).

The positivity can be really critical in geodynamical simulations: a negative density makes no sense, and a negative viscosity even less!

Example 3 The 2nd-order Bernstein polynomials are used in STONE 64. The actual context of this stone is not important. Fields such as density and viscosity are known on the 9 nodes of the Q_2 element and need to be projected onto the 9 quadrature points. For instance, these nodal fields are given by:
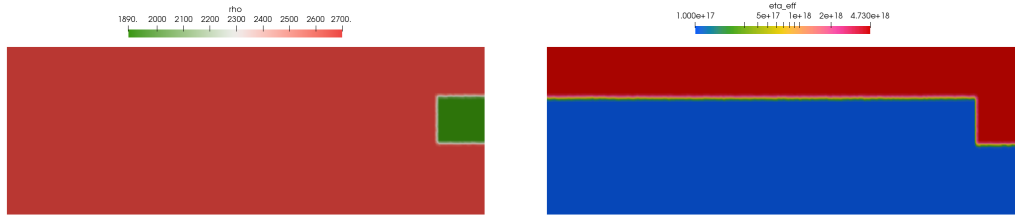
The resulting fields on the quadrature points are shown:
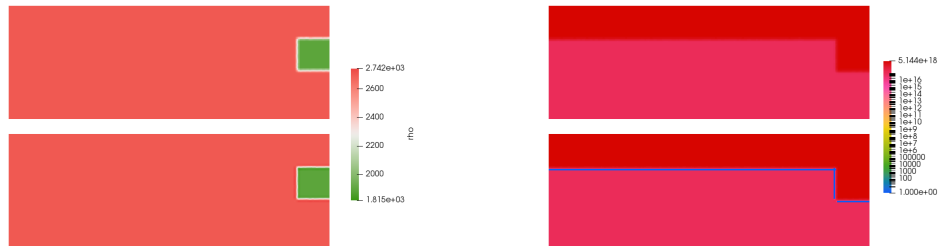
The bottom row is obtained with the basis functions while the top row is obtained with the Bernstein polynomials as interpolants. The thin blue line actually indicated points with negative viscosity and on the left the colour bar shows densities below the value of 1890 (lowest density in the domain).

Example 4 In this example I have created a small python code (in `/images/bernstein`) which creates a single Q_2 element in 2D, i.e. it consists of 9 nodes. 9 different scenarios are created by assigning some of the nodes either a density value of 1 or 0. The internal numbering is as follows:

```
3--6--2
|     |
7  8  5
|     |
0--4--1
```

I then proceed to generate 100,000 random points inside the element and I use either the Q_2 basis functions to interpolate the density onto them or the 2nd order Bernstein polynomials. The basis functions used are as follows:

```python
def NNN(r,s):
    val = np.zeros(9,dtype=np.float64)
    val[0]= 0.5*r*(r-1.)  *  0.5*s*(s-1.)
    val[1]= 0.5*r*(r+1.)  *  0.5*s*(s-1.)
    val[2]= 0.5*r*(r+1.)  *  0.5*s*(s+1.)
    val[3]= 0.5*r*(r-1.)  *  0.5*s*(s+1.)
    val[4]=     (1.-r**2) *  0.5*s*(s-1.)
    val[5]= 0.5*r*(r+1.)  *      (1.-s**2)
    val[6]=     (1.-r**2) *  0.5*s*(s+1.)
    val[7]= 0.5*r*(r-1.)  *      (1.-s**2)
    val[8]=     (1.-r**2) *      (1.-s**2)
```

```
    return val

def BBB(r,s):
    val = np.zeros(9,dtype=np.float64)
    val[0]=  0.25*(1-r)**2   *  0.25*(1-s)**2
    val[1]=  0.25*(1+r)**2   *  0.25*(1-s)**2
    val[2]=  0.25*(1+r)**2   *  0.25*(1+s)**2
    val[3]=  0.25*(1-r)**2   *  0.25*(1+s)**2
    val[4]=  0.5*(1-r**2)    *  0.25*(1-s)**2
    val[5]=  0.25*(1+r)**2   *  0.5*(1-s**2)
    val[6]=  0.5*(1-r**2)    *  0.25*(1+s)**2
    val[7]=  0.25*(1-r)**2   *  0.5*(1-s**2)
    val[8]=  0.5*(1-r**2)    *  0.5*(1-s**2)
    return val
```

In the following table I report the min/max values of the density on the swarm of points for both methods and all 9 cases:

case	$\min(\rho^h_{Q_2})$	$\max(\rho^h_{Q_2})$	$\min(\rho^h_{B_2})$	$\max(\rho^h_{B_2})$
1	-0.124978	0.994484	0.000000	0.996320
2	-0.124999	0.998679	0.000000	0.499537
3	-0.140625	1.124868	0.000000	0.999317
4	-0.125000	0.999997	0.000000	0.999998
5	-0.187699	1.124808	0.000000	0.999978
6	0.000000	1.265620	0.000000	0.999980
7	-0.225000	1.124911	0.000000	0.999996
8	-0.124418	1.224999	0.000000	1.000000
9	1.000000	1.000000	1.000000	1.000000

We find that using Q_2 basis functions yields values either above 1 or less than 0, which makes little sense, while the B_2 functions produce values exactly between 0 and 1. Only case 2 with B_2 produces values up to 0.5, which is understandable since the function $B_{2,7}$ (i.e. at node 7) reaches a maximum value of 1/2.

Also, the last case 9 corresponds to assigning all nodes $\rho = 1$, so that we expect the min/max values to be exactly 1 (sanity check).

I show below from left to right: the nodal values, the Q_2-based field, where it is negative, and the B_2-based field for all nine cases:

- Case 1:
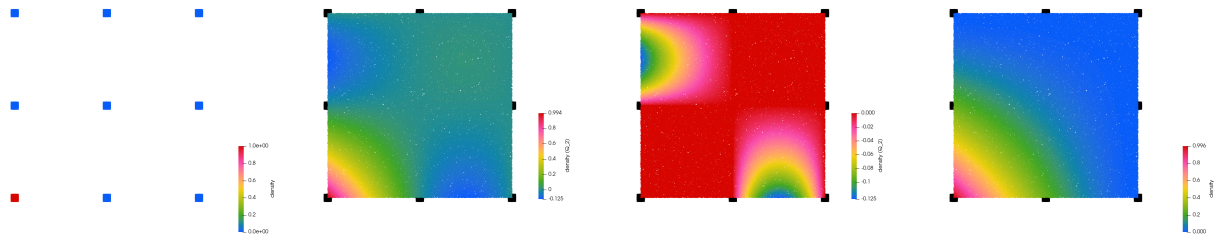
- Case 2:
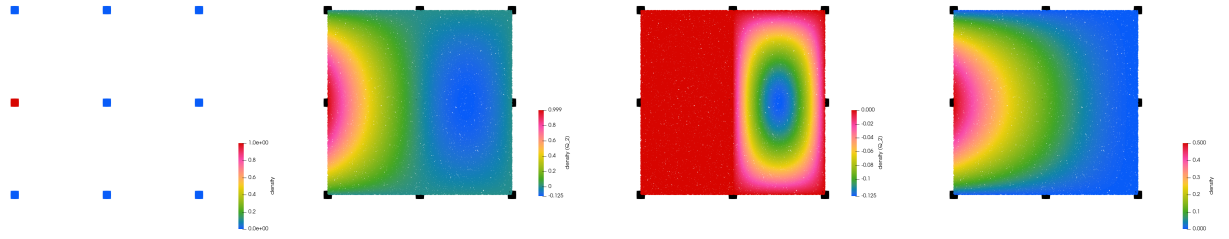
- Case 3:
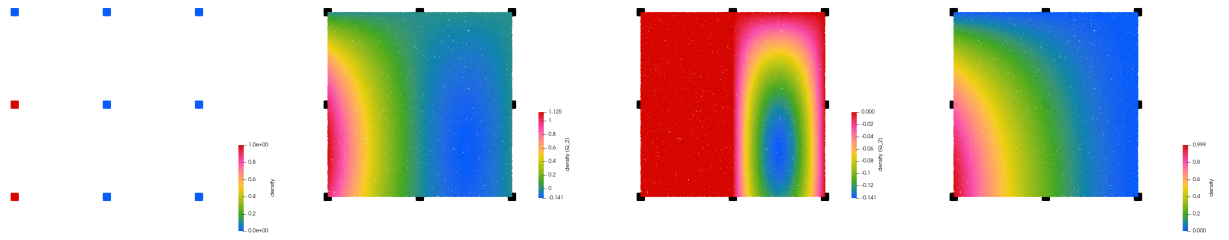
- Case 4:
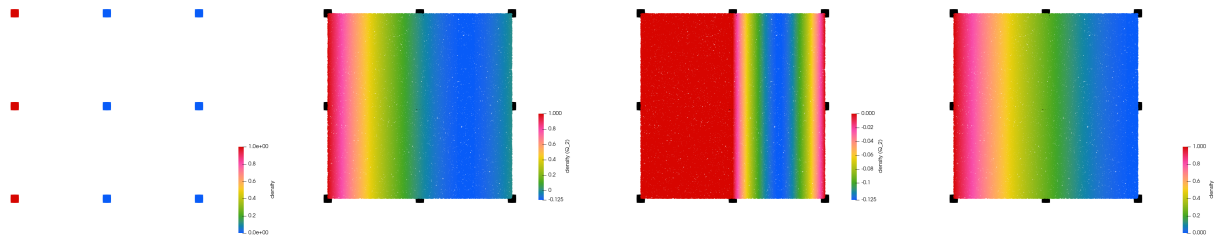
- Case 5:
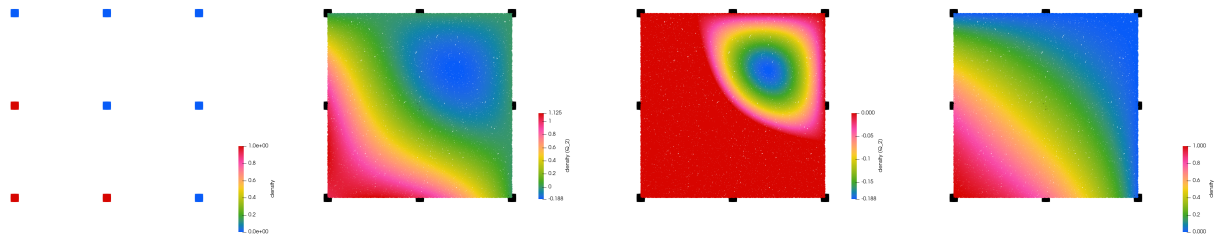
- Case 6:
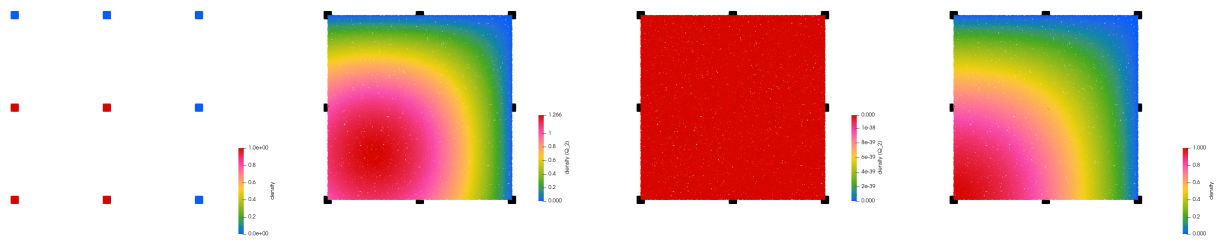
- Case 7:
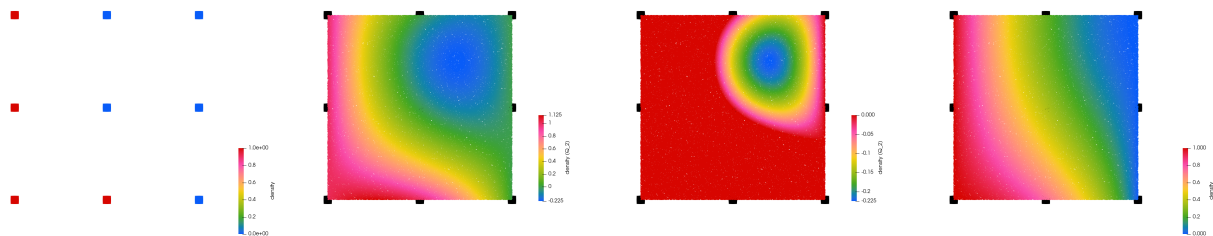
- Case 8:
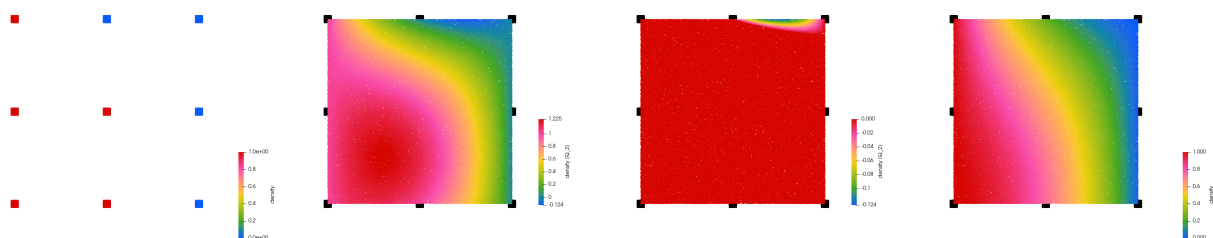
- Case 9:
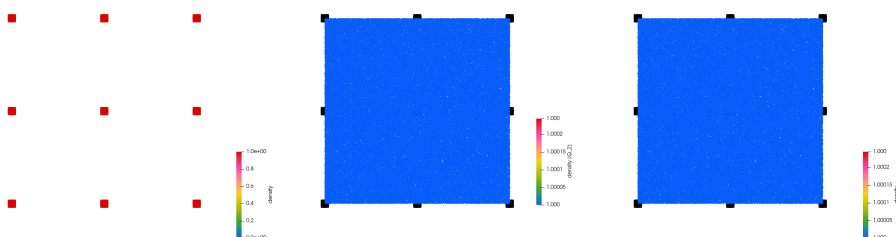
16.27 Computing field derivatives - see overleaf

currently being worked on in Overleaf

16.28 Iterative solvers

In what follows, we want to solve the system of linear equations

$$\boldsymbol{A} \cdot \vec{x} = \vec{b} \tag{16.109}$$

for the vector $\vec{x}$. We denote the unique solution of this system by $\vec{x}^{\star}$.

Note that in some cases the the known $n \times n$ matrix $\boldsymbol{A}$ is symmetric (i.e., $\boldsymbol{A}^T = \boldsymbol{A}$), positive-definite (i.e. $\vec{x}^T \cdot \boldsymbol{A} \cdot \vec{x} > 0$ for all non-zero vectors $\vec{x}$ in $\mathbb{R}^n$), and real, and $\vec{b}$ is known as well (typically the $\mathbb{K}$ matrix).

📖 Relevant Literature: Direct and Iterative Solvers [788]

Stationary iterative methods

Basic examples of stationary iterative methods use a splitting of the matrix $\boldsymbol{A}$ such as

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}$$

where $\boldsymbol{D}$ is only the diagonal part of $\boldsymbol{A}$, $\boldsymbol{L}$ is the strict lower triangular part of $\boldsymbol{A}$ and $\boldsymbol{U}$ is the strict upper triangular part of $\boldsymbol{A}$.

For instance:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 5 & 8 \\ 6 & 4 & 2 \\ -1 & 7 & 5 \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{pmatrix} \quad \boldsymbol{L} = \begin{pmatrix} 0 & 0 & 0 \\ 6 & 0 & 0 \\ -1 & 7 & 0 \end{pmatrix} \quad \boldsymbol{U} = \begin{pmatrix} 0 & 5 & 8 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

- Jacobi method[35]: The solution is then obtained iteratively via

$$\boldsymbol{D} \cdot \vec{x}^{k+1} = -(\boldsymbol{L} + \boldsymbol{U}) \cdot \vec{x}^k + \vec{b} \qquad k = 0, 1, \ldots \tag{16.110}$$

  where $\vec{x}^k$ is the k-th approximation or iteration of $\vec{x}$ and $\vec{x}^0$ is the initial guess (often taken to be zero). A sufficient (but not necessary) condition for the method to converge is that the matrix $\boldsymbol{A}$ is strictly or irreducibly diagonally dominant.

- Gauss-Seidel method[36]: It is defined by the iteration

$$\boldsymbol{L}_{\star} \cdot \vec{x}^{k+1} = -\boldsymbol{U} \cdot \vec{x}^k + \vec{b} \qquad k = 0, 1, \ldots \tag{16.111}$$

  where $\boldsymbol{L}_{\star} = \boldsymbol{L} + \boldsymbol{U}$.

  There is actually a way to make the computation of $\vec{x}^{(k+1)}$ which uses the elements of $\vec{x}^{(k+1)}$ that have already been computed, and only the elements of $\vec{x}^{(k)}$ that have not been computed in the $k + 1$ iteration. This means that, unlike the Jacobi method, only one storage vector is required as elements can be overwritten as they are computed, which can be advantageous for very large problems.

  Note that Gauss-Seidel is the same as SOR (successive over-relaxation) with $\omega = 1$.

- Successive over-relaxation method (SOR):

$$(\boldsymbol{D} + \omega \boldsymbol{L}) \cdot \vec{x}^{k+1} = -(\omega \boldsymbol{U} + (\omega - 1)\boldsymbol{D}) \cdot \vec{x}^k + \omega \vec{b} \qquad k = 0, 1, \ldots \tag{16.112}$$

---

[35]https://en.wikipedia.org/wiki/Jacobi_method
[36]https://en.wikipedia.org/wiki/Gauss-Seidel_method

- Symmetric successive over-relaxation (SSOR)[37]: The version of SOR for symmetric matrices $\boldsymbol{A}$, in which $\boldsymbol{U} = \boldsymbol{L}^T$ is given by the recursion

$$\vec{x}^{k+1} = \vec{x}^k - \gamma^k \boldsymbol{P}^{-1}(\boldsymbol{A} \cdot \vec{x}^{(k)} - \vec{b}) \qquad k = 0, 1, \ldots \tag{16.113}$$

with

$$\boldsymbol{P} = \left(\frac{\boldsymbol{D}}{\omega} + \boldsymbol{L}\right) \frac{\omega}{2 - \omega} \boldsymbol{D}^{-1} \cdot \left(\frac{\boldsymbol{D}}{\omega} + \boldsymbol{L}\right)$$

with $0 < \omega < 2$.

All these methods can be cast in a more general framework[38]: The basic iterative methods work by splitting the matrix A into M-N and here the matrix M should be easily invertible. The iterative methods are now defined as

$$\boldsymbol{M} \cdot \vec{x}^{k+1} = \boldsymbol{N} \cdot \vec{x}^k + \vec{b} \tag{16.114}$$

with

- Richardson method: $\boldsymbol{M} = \frac{1}{\omega}\boldsymbol{I}$

- Jacobi method: $\boldsymbol{M} = \boldsymbol{D}$

- Damped Jacobi method: $\boldsymbol{M} = \frac{1}{\omega}\boldsymbol{D}$

- Gauss-Seidel method: $\boldsymbol{M} = \boldsymbol{D} + \boldsymbol{L}$

- Successive over-relaxation method: $\boldsymbol{M} = \frac{\boldsymbol{D}}{\omega} + \boldsymbol{L}$

- Symmetric successive over-relaxation: $\boldsymbol{M} = \left(\frac{\boldsymbol{D}}{\omega} + \boldsymbol{L}\right) \frac{\omega}{2-\omega} \boldsymbol{D}^{-1} \cdot \left(\frac{\boldsymbol{D}}{\omega} + \boldsymbol{L}\right)$

and $\boldsymbol{N} = \boldsymbol{M} - \boldsymbol{A}$.

Krylov subspace methods

- Conjugate Gradient [39]

  It was first proposed by Hestenes and Stiefel in 1952 [600]. The method solves an SPD system $\boldsymbol{A} \cdot \vec{x} = \vec{b}$ of size n. In theory (i.e. exact arithmetic) it does so in n iterations. Each iteration requires a few inner products in $\mathbb{R}^n$ and one matrix-vector multiplication. With roundoff error, CG can work poorly (or not at all), but for some $\boldsymbol{A}$ (and $\vec{b}$), can get good approximate solution in $\ll n$ iterations.

  As an iterative method, the conjugate gradient method monotonically (in the energy norm) improves approximations $\vec{x}_k$ to the exact solution and may reach the required tolerance after a relatively small (compared to the problem size) number of iterations. The improvement is typically linear and its speed is determined by the condition number $\kappa(\boldsymbol{A})$ of the system matrix $\boldsymbol{A}$: the larger $\kappa(\boldsymbol{A})$ is, the slower the improvement.

  If $\kappa(\boldsymbol{A})$ is large, preconditioning is commonly used to replace the original system $\boldsymbol{A} \cdot \vec{x} - \vec{b} = \vec{0}$ with $\boldsymbol{M}^{-1} \cdot (\boldsymbol{A} \cdot \vec{x} - \vec{b}) = \vec{0}$ such that $\kappa(\boldsymbol{M}^{-1} \cdot \boldsymbol{A})$ is smaller than $\kappa(\boldsymbol{A})$.

  The resulting method is called the Preconditioned Conjugate Gradient method (PCG). An extreme case of preconditioner is $\boldsymbol{M} = \boldsymbol{A}^{-1}$ but it is a silly case since applying the preconditioner is as difficult as solving the system in the first place. In the end the goal is to find a matrix $\boldsymbol{M}$ that is cheap to multiply, and is an approximate inverse of $\boldsymbol{A}$ (or at least has a more clustered spectrum than $\boldsymbol{A}$).

---

[37]https://en.wikipedia.org/wiki/Symmetric_successive_over-relaxation
[38]https://en.wikipedia.org/wiki/Iterative_method
[39]https://en.wikipedia.org/wiki/Conjugate_gradient_method

Left (CG):

$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$
if $\mathbf{r}_0$ is sufficiently small, then return $\mathbf{x}_0$ as the result
$\mathbf{p}_0 := \mathbf{r}_0$
$k := 0$
repeat
$$\alpha_k := \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}$$
$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$
$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{p}_k$
if $\mathbf{r}_{k+1}$ is sufficiently small, then exit loop
$$\beta_k := \frac{\mathbf{r}_{k+1}^\top \mathbf{r}_{k+1}}{\mathbf{r}_k^\top \mathbf{r}_k}$$
$\mathbf{p}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$
$k := k + 1$
end repeat
return $\mathbf{x}_{k+1}$ as the result

Right (PCG):

$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{z}_0 := \mathbf{M}^{-1}\mathbf{r}_0$
$\mathbf{p}_0 := \mathbf{z}_0$
$k := 0$
repeat
$$\alpha_k := \frac{\mathbf{r}_k^\top \mathbf{z}_k}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}$$
$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$
$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{p}_k$
if $\mathbf{r}_{k+1}$ is sufficiently small then exit loop end if
$\mathbf{z}_{k+1} := \mathbf{M}^{-1}\mathbf{r}_{k+1}$
$$\beta_k := \frac{\mathbf{z}_{k+1}^\top \mathbf{r}_{k+1}}{\mathbf{z}_k^\top \mathbf{r}_k}$$
$\mathbf{p}_{k+1} := \mathbf{z}_{k+1} + \beta_k \mathbf{p}_k$
$k := k + 1$
end repeat
The result is $\mathbf{x}_{k+1}$

Let's put it all together into one piece now. The method of Conjugate Gradients is:

$$d_{(0)} = r_{(0)} = b - Ax_{(0)},$$

$$\alpha_{(i)} = \frac{r_{(i)}^T r_{(i)}}{d_{(i)}^T A d_{(i)}} \qquad \text{(by Equations 32 and 42)},$$

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} d_{(i)},$$

$$r_{(i+1)} = r_{(i)} - \alpha_{(i)} A d_{(i)},$$

$$\beta_{(i+1)} = \frac{r_{(i+1)}^T r_{(i+1)}}{r_{(i)}^T r_{(i)}},$$

$$d_{(i+1)} = r_{(i+1)} + \beta_{(i+1)} d_{(i)}.$$

Top: algorithms as obtained from Wikipedia (Left: CG; Right: PCG); Bottom: algorithm from Shewchuk (1994) [1234].

Also available on Wikipedia is a (naive) MATLAB implementation of the CG algorithm:

```matlab
function x = conjgrad(A, b, x)
    r = b - A * x;
    p = r;
    rsold = r' * r;

    for i = 1:length(b)
        Ap = A * p;
        alpha = rsold / (p' * Ap);
        x = x + alpha * p;
        r = r - alpha * Ap;
        rsnew = r' * r;
        if sqrt(rsnew) < 1e-10
                break
        end
        p = r + (rsnew / rsold) * p;
        rsold = rsnew;
    end
end
```

We see that its implementation is actually rather simple and straightforward!

Relevant Literature: Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain [1234]. CG using mpi [1243]. Een kwart eeuw iteratieve methoden [1431].

The CG and PCG algorithms are used in Section 10.13.3. It is implemented in STONE 15,16,82.

- Biconjugate Gradient method [40]

- Biconjugate Gradient stabilised method [41]

---

[40] https://en.wikipedia.org/wiki/Biconjugate_gradient_method
[41] https://en.wikipedia.org/wiki/Biconjugate_gradient_stabilized_method

- MINRES: For iterative solution of symmetric systems Ax = b, the conjugate gradient method (CG) is commonly used when A is positive definite, while the minimum residual method (MINRES) is typically reserved for indefinite systems.

- Generalized minimal residual method (GMRES) [42]

---

[42] https://en.wikipedia.org/wiki/Generalized_minimal_residual_method

16.29 Weak seeds in extension modelling

This section was mostly written by I. van Zelst with some input by S. Buiter.

Numerical models that investigate dynamics of the lithosphere and upper mantle always start from an initial geometry with a set of prescribed mechanical and thermal conditions. This initial setup is usually a more-or-less standard representation of the lithosphere and asthenosphere, as defined from compilations of geological and geophysical observations and laboratory measurements. Deformation is driven by internal buoyancy forces and/or velocity or stress boundary conditions. However, unless an intrinsically unstable setup is defined or boundary conditions are discontinuous, deformation may take long model time to localize (up to millions of years). This is because these models need to build up numerical disturbance to create starting points for the deformation. In such models, deformation may in the first stages be accommodated by pure shear extension or shortening [1089, 954].

To avoid this long starting phase and, in addition, exert some control over the initial location of deformation (preferably away from the boundaries), modelers use different approaches to initiate and localize deformation.

One manner to localize deformation is by discontinuous boundary conditions, such as the so-called S-point velocity discontinuity at the bottom of the system (or the tip of a basal sheet) which is used in both numerical [151, 391, 1460, 66, 175, 1347, 145] and analogue studies [175, 921]. These models are usually on the scale of the (upper-) crust. S-point models are less flexible than upper-mantle scale models as they do not include feedback relations between deformation and the basal velocity field. Models of extension of continental lithosphere often use 'seeds' to initiate extension. Such seeds are usually small regions that are weaker than the surrounding crust and lithosphere. The use of seeds can be justified by considering the fact that in nature continental lithosphere is hardly ever (if at all) homogeneous in composition and stratification. In addition, extension often occurs in regions of former convergence, such as the opening of the North Atlantic Ocean that largely followed the old sutures of the Iapetus and Rheic Oceans [1464]. Analogues for numerical seeds can therefore be found in inherited faults, inherited crustal thickness changes, and/or plumes impacting the lithosphere. However, this immediately points out a problem with single-seed models as orogenic inheritance and mantle upwellings may be expected to occur over larger areas than a seed of some hundreds of meters to a few kilometers in width and height.

A literature survey shows that seeds in previous numerical studies differ in shape, size, orientation, mechanical and thermal properties, and depth in the models. Three types of weak seeds can be identified that have been used in previous models of (continental) extension:

Seeding through thermal effects   A weak region can be achieved by a temperature anomaly in the crust or lithosphere [191], which is created by directly imposing a temperature difference, by assigning high radiogenic heat production, or modeling a thermal upwelling in the mantle below. The elevated temperature reduces viscosity values for models with a temperature-dependent viscosity. An advantage of using an imposed temperature anomaly is that it will dissipate with time, thus reducing the impact on later model stages [565]. Examples of thermal anomalies used to initiate extension are an elevated temperature at the base of the crust [565], an elevated temperature at the base of the lithosphere ($100°$ in [198], up to $200°$ in [169]), a temperature anomaly imposed from the base of the lithosphere to the middle crust [250], and a $10\mathrm{mW\,m^{-2}}$ perturbation in basal heat flow [442]. In [198] the rifting is initiated by means of a thermal perturbation placed at the bottom of the mantle lithosphere with a maximum temperature T_2 exponentially which decays from the center to T_1 on the left and T_3 on the right.

Seeding by mechanical inhomogeneity A seed may be composed of a material with a lower rheological strength than the surroundings. A weak seed may, for example, have a lower imposed viscosity [810, 720, 927], a lower value for angle of internal friction [1088, 723, 1343, 512, 245], a lower value for cohesion [10], or a lower value for density [1358].

The seed may also be assigned different material properties, as, for example, a Von Mises seed in a frictional plastic material [647]. A frequently used approach is to assume that a region has already accumulated strain, leading to strain-weakening [799, 648, 1057, 10, 9, 754, 11]. Previous studies have used a variety of shapes and sizes for weak seeds. Examples are square seeds, fault-shaped weak inclusions, and rectangular seeds with different aspect ratios: [648] use a 6×3 km seed, while the weak seed of [647] has a size of 12×10 km.

Instead on confining the seed to a geometrically simple region, randomly distributed seeds have also been used [1343], [1348], albeit for compression.

Seeding through geometrical discontinuity A seed is created by an abrupt variation in the thickness of the crust and/or lithosphere. A locally thinned crust could be thought to be caused by a previous rifting phase, whereas a thicker crust could represent preceding mountain building. Such crustal thickness variations effect not only mechanical strength, but may also impose a thermal anomaly. Burg & Schmalholz [192] implemented a Gaussian shaped mohorovičić discontinuity of 250m height as a representation of the weak zone resulting in a slightly thinner crust. A step change in crustal thickness alters the symmetry of the domain. Chenin *et al.* [248] implement a sinusoidal perturbation of the Moho.

Only few studies have investigated how different methods of implementing a weak zone can affect the results of a model. [379] found that a single seed produces a symmetric narrow rift, an initial shear zone tends to produce an asymmetric rift, and multiple seeds promote a wide rift. Note however that this behavior will be affected by rheological stratification, as not all systems can evolve in a wide rift mode [648, 177]. [379] found that a seed needs to be 10 times weaker than the surrounding material in order to localize strain. To initiate shear bands with a Coulomb dip angle ($45 \pm \phi/2$, where ϕ is the angle of internal friction), seeds need to be well resolved (5-10 elements, [720]).

This variety in the shape, size, orientation, mechanical and thermal properties, and depth of the seed(s) begs the question if these different approaches to initiate extension could have an effect on model evolution? As such variations might not be removed by subsequent deformation stages, the initiation effects could propagate into later model evolution. In addition, weak seeds introduce a weakness into the extensional system that may potentially be long-lasting. For instance, seeds with a weakness defined by material or strain-weakened properties stay in the model and may control deformation also in later model stages. The heat associated with thermal seeds will diffuse away, but additional heat has been introduced into the initial system and the setup will therefore differ from models with mechanically weak seeds.

16.30 Computing the volume of a hexahedron

What follows is based on the report "Efficient Computation of Volume of Hexahedral Cells" by J. Grandy (1997) [510]. We assume the following internal numbering of the hexahedron, which is different than the one in the paper:
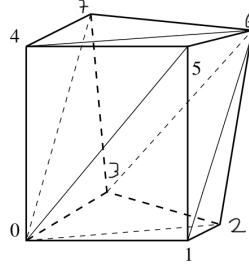
Modified from [510]

If the hexahedron is such that some or all the opposite faces are planes parallel to each other than the volume can be arrived at very simply[43]. The real catch here is that the four nodes which make up a face are not necessarily co-planar!

The volume can then be computed as follows

$$
\begin{aligned}
V \;=\; & [(\vec{r}_6 - \vec{r}_1) + (\vec{r}_7 - \vec{r}_0), (\vec{r}_6 - \vec{r}_3), (\vec{r}_2 - \vec{r}_0)] \\
+\; & [(\vec{r}_7 - \vec{r}_0), (\vec{r}_6 - \vec{r}_3) + (\vec{r}_5 - \vec{r}_0), (\vec{r}_6 - \vec{r}_4)] \\
+\; & [(\vec{r}_6 - \vec{r}_1), (\vec{r}_5 - \vec{r}_0), (\vec{r}_6 - \vec{r}_4) + (\vec{r}_2 - \vec{r}_0)] \\
/\; & 12
\end{aligned}
\tag{16.115}
$$

where $[\cdot]$ is the triple product:

$$
[\vec{A}, \vec{B}, \vec{C}] = \begin{vmatrix} A_x & B_x & C_x \\ A_y & B_y & C_y \\ A_z & B_z & C_z \end{vmatrix}
$$

It is implemented and used in STONE 98. The code is shown hereunder:

```
def hexahedron_volume (x,y,z):
    val=(triple_product(x[6]-x[1]+x[7]-x[0],y[6]-y[1]+y[7]-y[0],z[6]-z[1]+z[7]-z[0],\
                        x[6]-x[3],              y[6]-y[3],             z[6]-z[3],            \
                        x[2]-x[0],              y[2]-y[0],             z[2]-z[0]           )\
        +triple_product(x[7]-x[0],              y[7]-y[0],             z[7]-z[0],            \
                        x[6]-x[3]+x[5]-x[0],y[6]-y[3]+y[5]-y[0],z[6]-z[3]+z[5]-z[0],\
                        x[6]-x[4],              y[6]-y[4],             z[6]-z[4]           )\
        +triple_product(x[6]-x[1],              y[6]-y[1],             z[6]-z[1],            \
                        x[5]-x[0],              y[5]-y[0],             z[5]-z[0],            \
                        x[6]-x[4]+x[2]-x[0],y[6]-y[4]+y[2]-y[0],z[6]-z[4]+z[2]-z[0]))
    /12.
    return val
```

with

```
def triple_product (Ax,Ay,Az,Bx,By,Bz,Cx,Cy,Cz):
    val = Ax * ( By * Cz - Bz * Cy )\
        - Ay * ( Bx * Cz - Bz * Cx )\
        + Az * ( Bx * Cy - By * Cx )
    return val
```

---

[43]https://en.wikipedia.org/wiki/Cuboid

📚 Relevant Literature: J.K. Dukowicz. "Efficient Volume Computation for Three- Dimensional Hexahedral Cells". In: *J. Comp. Phys.* 74 (1988), pp. 493–496. DOI: `10.1016/0021-9991(88) 90091-5`

16.31 Bandwidth reduction, matrix reordering

The need for reordering

The profile (or envelope) of a symmetric matrix determines how close its non-zero elements are to the diagonal:

$$\text{profile} = \sum_{i=1}^{n}(i - \min(ne(i)))$$

The bandwidth is the largest deviation:

$$\text{bandwidth} = \max(i - \min(ne(i)))$$

The cost for a band cholesky factorisation with bandwidth p is $n(p^2 + 3p)$ flops assuming $n >> p$. (source?)

In conclusion: reducing bandwidth means a factor solve if Cholesky factorisation is used.

A simple example

Let us consider a structurally symmetric matrix $\boldsymbol{M}$. We wish to reduce its bandwidth by permuting rows and columns such as to move all the nonzero elements of $\boldsymbol{M}$ in a band as close as possible to the diagonal. We then talk about *Bandwidth Reduction*.

We know that the solution a linear system remains unchanged if lines or columns of the matrix (and corresponding rhs) are permuted.

For example[44] , let us consider the 5×5 matrix $\boldsymbol{M}$:

$$\boldsymbol{M} = \begin{pmatrix} 1 & . & . & . & 1 & . & . & . \\ . & 1 & 1 & . & . & 1 & . & 1 \\ . & 1 & 1 & . & 1 & . & . & . \\ . & . & . & 1 & . & . & 1 & . \\ 1 & . & 1 & . & 1 & . & . & . \\ . & 1 & . & . & . & 1 & . & 1 \\ . & . & . & 1 & . & . & 1 & . \\ . & 1 & . & . & . & 1 & . & 1 \end{pmatrix}$$

Simply through row and column permutations it can be rewritten

$$\boldsymbol{M}' = \begin{pmatrix} 1 & 1 & . & . & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . \\ . & . & 1 & 1 & 1 & . & . & . \\ . & . & 1 & 1 & 1 & . & . & . \\ . & . & 1 & 1 & 1 & 1 & . & . \\ . & . & . & . & 1 & 1 & 1 & . \\ . & . & . & . & . & 1 & 1 & 1 \\ . & . & . & . & . & . & 1 & 1 \end{pmatrix}$$

The different existing algorithms

- The simplest bandwidth reduction method is the Cuthill-McKee algorithm (1969) [316]

- Reverse Cuthill-McKee algorithm (1976) [488] - see STONE 78,82

---

[44]Taken from `http://ciprian-zavoianu.blogspot.com/2009/01/project-bandwidth-reduction.html`

- The Gibbs-Poole-Stockmeyer and Gibbs-King algorithm is an alternative, often superior, profile reduction method (1976) [489]

- Sloan algorithm [1253, 1252]

Implementation in python

The documentation for the reverse Cuthill-McKee algorithm is available online[45], but it is quite useless as to how the result of the function call should be used. I therefore provide here a small python program in /images/reordering which builds matrix M and returns M'.

This header is necessary:

```
from scipy.sparse import csr_matrix
from scipy.sparse.csgraph import reverse_cuthill_mckee
```

After the matrix is filled, the reverse Cuthill-McKee algorithm is used to compute the permutation order of the rows and columns:

```
perm = reverse_cuthill_mckee(sparse_matrix,symmetric_mode=True)
```

The result is an array of size n which indicates the new order of rows and columns:

```
[6 3 7 5 1 2 4 0]
```

All that we have to do then is to use this array to rebuild the matrix:

```
sparse_matrix=sparse_matrix[np.ix_(perm,perm)]
```

In case a right hand side vector exists, it must be reordered too in order to match the matrix:

```
rhs=rhs[np.ix_(perm)]
```

Assuming a solution of the reordered matrix and rhs has been obtained, it must be reordered back. We therefore create the inverse permutation array:

```
perm_inv=np.empty(n,dtype=np.int32)
for i in range(0,n):
    perm_inv[perm[i]]=i
```

and use it as follows:

```
sol=sol[np.ix_(perm_inv)]
```


a) before reordering; b) after reordering.
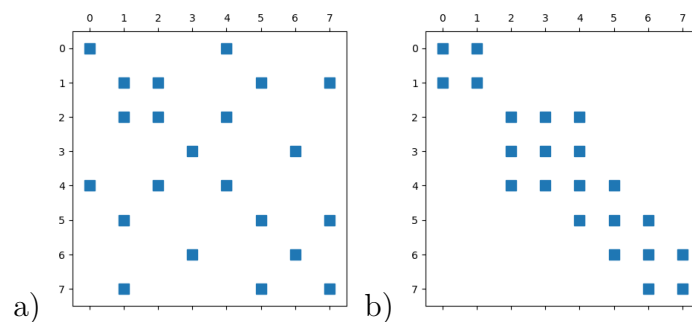
[45]https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.reverse_cuthill_mckee.html

16.32 Scaling between dimensioned and dimensionless quantities

All quantities with an r subscript denote real quantities, while quantities with an m subscript denote dimensionless model quantites, i.e. values that are used as inputs to the code.

Let us define four coefficients K_L, K_T, K_M and K_θ as follows:

$$K_L = \frac{L_m}{L_r} \qquad K_T = \frac{T_m}{T_r} \qquad K_M = \frac{M_m}{M_r} \qquad K_\theta = \frac{\theta_m}{\theta_r} \qquad (16.116)$$

where $L_{\{m,r\}}$, $T_{\{m,r\}}$, $M_{\{m,r\}}$ and $\theta_{\{m,r\}}$ are respectively lengths, times, masses and temperatures, and the coefficients K bear the following dimensions:

$$K_M \to kg^{-1} \text{ , or } M^{-1}$$

$$K_L \to m^{-1} \text{ , or } L^{-1}$$

$$K_T \to s^{-1} \text{ , or } T^{-1}$$

$$K_\theta \to K^{-1} \text{ , or } \theta^{-1}$$

Velocity The dimensions of a velocity are: $m.s^{-1}$, or LT^{-1}, so that

$$\boldsymbol{v}_m = \frac{L_m}{T_m} = \frac{K_L L_r}{K_T T_r} = K_L K_T^{-1} \frac{L_r}{T_r} = K_L K_T^{-1} \boldsymbol{v}_r \qquad (16.117)$$

Viscosity The dimensions of a viscosity are $Pa.s$, or $ML^{-1}T^{-1}$, so that

$$\mu_m = \frac{M_m}{L_m T_m} = \frac{K_M}{K_L K_T} \frac{M_r}{L_r T_r} = \frac{K_M}{K_L K_T} \mu_r \qquad (16.118)$$

Cohesion The dimensions of a cohesion are those of a stress, i.e. Pa, or $ML^{-1}T^{-2}$, so that

$$c_m = \frac{M_m}{L_m T_m^2} = \frac{K_M}{K_L K_T^2} \frac{M_r}{L_r T_r^2} = \frac{K_M}{K_L K_T^2} c_r$$

Density The dimensions of a density are $kg.m^{-3}$, or ML^{-3}, so that

$$\rho_m = \frac{M_m}{L_m^3} = \frac{K_M}{K_L^3} \frac{M_r}{L_r^3} = \frac{K_M}{K_L^3} \rho_r$$

Gravity The dimensions of the gravity acceleration are $m.s^{-2}$, or LT^{-2}, so that

$$\boldsymbol{g}_m = \frac{L_m}{T_m^2} = \frac{K_L}{K_T^2} \frac{L_r}{T_r^2} = \frac{K_L}{K_T^2} \boldsymbol{g}_r$$

Body force The dimensions of body forces are those of $\rho\boldsymbol{g}$, i.e. $kg.m^{-2}.s^{-2}$, or $ML^{-2}T^{-2}$, so that

$$(\rho\boldsymbol{g})_m = \frac{M_m}{L_m^2 T_m^2} = \frac{K_M}{K_L^2 K_T^2} \frac{M_r}{L_r^2 T_r^2} = \frac{K_M}{K_L^2 K_T^2} (\rho\boldsymbol{g})_r$$

Thermal expansion The dimension of thermal expansion is $°C^{-1}$, so that

$$\alpha_m = \frac{1}{\theta_m} = \frac{1}{K_\theta} \frac{1}{\theta_r} = \frac{1}{K_\theta} \alpha_r$$

Thermal conductivity   The dimensions of the thermal conductivity are $W/m/K$, i.e. $kg.m.s^{-3}.K^{-1}$, so that

$$k_m = \frac{M_m L_m}{T_m^3 \theta_m} = \frac{K_M K_L}{K_T^3 K_\theta} \frac{M_r L_r}{T_r^3 \theta_r} = \frac{K_M K_L}{K_T^3 K_\theta} k_r$$

Thermal diffusivity   The dimensions of thermal diffusivity are m^2/s so that

$$\kappa_m = \frac{L_m^2}{T_m} = \frac{K_L^2}{K_T} \frac{L_r^2}{T_r} = \frac{K_L^2}{K_T} \kappa_r$$

Heat capacity   The dimensions are $J.kg^{-1}.K^{-1}$, i.e. $m^2.s^{-2}.K^{-1}$ so that

$$(c_p)_m = \frac{L_m^2}{T_m^2 \theta_m} = \frac{K_L^2}{K_T^2 K_\theta} \frac{L_r^2}{T_r^2 \theta_r} = \frac{K_L^2}{K_T^2 K_\theta} (c_p)_r$$

Radiogenic heat production   The dimensions are $W.m^{-3}$, i.e. $kg.m^{-1}.s^{-3}$ so that

$$H_m = \frac{M_m}{L_m T_m^3} = \frac{K_M}{K_L K_T^3} \frac{M_r}{L_r T_r^3} = \frac{K_M}{K_L K_T^3} H_r$$

A constant   The dimensions of A are $Pa^{-n}.s^{-1}$, or $M^{-n} L^n T^{(2n-1)}$ so that

$$A_m = \frac{L_m^n T_m^{2n+1}}{M_m^n} = \frac{K_L^n K_T^{2n-1}}{K_M^n} A_r$$

Activation energy   The dimensions of Q are $J.mol^{-1}$, but the dimensions of $\tilde{Q} = Q/R$ are K, so that

$$\tilde{Q}_m = \theta_m = K_\theta \tilde{Q}_r$$

Activation volume   The dimensions of V are $m^3.mol^{-1}$, but the dimensions of $\tilde{V} = V/R$ are $m^3.°C.J^{-1}$, so that

$$\tilde{V}_m = \frac{K_L K_T^2 K_\theta}{K_M} \tilde{V}_r$$

The following coefficients are used in the code too :

$$K_\mu = \frac{\mu_m}{\mu_m} = \frac{K_M}{K_L K_T}$$

$$K_v = \frac{\boldsymbol{v}_m}{\boldsymbol{v}_r} = \frac{K_L}{K_T}$$

$$K_{stress} = \frac{K_M}{K_L K_T^2}$$

$$K_\rho = \frac{\rho_m}{\rho_r} = \frac{K_M}{K_L^3}$$

16.33 Spectral methods

Trubitsyn *et al.* (2008) [1376]

16.34 Spherical harmonics

We start from the Poisson equation for the gravity potential:

$$\Delta U = 4\pi \mathcal{G} \rho(\vec{r}) \tag{16.119}$$

As a consequence, inside a domain where $\rho = 0$, the equation becomes $\Delta U = 0$.

Let us assume that the spherical coordinates are appropriate for the problem at hand, and that the potential can be decomposed as follows:

$$U(r, \theta, \phi) = U_r(r) U\!\perp(\theta, \phi)$$

The full Laplacian operator in spherical coordinates is given by[46]:

$$\Delta U = \underbrace{\frac{1}{r^2} \frac{\partial}{\partial r}\left(r^2 \frac{\partial U}{\partial r} \right)}_{\Delta_r} + \underbrace{\frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial U}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2 U}{\partial \phi^2}}_{\Delta_\perp}$$

we then have:

$$(\Delta_r + \Delta_\perp)(U_r U_\perp) = 0$$

i.e.,

$$U_\perp \Delta_r U_r + U_r \Delta_\perp U_\perp = 0$$

Assuming $U_\perp = \sum_l \sum_m U_{lm} Y_{lm}$, knowing that spherical harmonics functions verify

$$r^2 \Delta_\perp Y_l^m(\theta, \phi) = -l(l+1) Y_l^m(\theta, \phi)$$

and assuming for now that the problem at hand is 1st degree (l=1), then

$$\Delta_\perp Y_l^m(\theta, \phi) = -\frac{2}{r^2} Y_l^m(\theta, \phi)$$

and then

$$\Delta_r U_r - U_r \frac{2}{r^2} = 0$$

make a link with my 2018 paper.

---

[46]https://en.wikipedia.org/wiki/Laplace_operator

In spherical coordinates, the Laplacian is given by

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}$$

We wish to solve Laplace's equation $\Delta T(r,\theta,\phi) = 0$ using the method of separation of variables:

$$T(r,\theta,\phi) = R(r)\Theta(\theta)\Phi(\phi)$$

We can insert this decomposition into the Laplace equation and multiply it by $r^2/R\Theta\Phi$ to obtain

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{1}{\Theta}\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{1}{\Phi}\frac{1}{\sin^2\theta}\frac{d^2\Phi}{d\phi^2} = 0$$

For reasons that will become clear later, the separation constant is taken to be $-m^2$:

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = -m^2 \tag{16.120}$$

$$-\frac{\sin\theta}{\Theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{\sin^2\theta}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -m^2 \tag{16.121}$$

The first equation yields

$$\Phi(\phi) = \begin{cases} e^{im\phi} \\ e^{-im\phi} \end{cases} \qquad \text{for } m = 0, 1, 2, 3, ...$$

Note that m must be an integer since ϕ is a periodic variable and $\Phi(\phi + 2\pi) = \Phi(\phi)$. In the case of $m = 0$, the general solution is $\Phi(\phi) = a\phi + b$, but we must choose $a = 0$ to be consistent with $\Phi(\phi + 2\pi) = \Phi(\phi)$. Hence in the case of $m = 0$, only one solution is allowed.

Eq. (16.121) can now be recast in the following form:

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -\frac{1}{\Theta}\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{m^2}{\sin^2\theta} \tag{16.122}$$

where the separation variable at this step is denoted by $l(l+1)$ for reasons that will shortly become clear. The resulting radial equation is

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = l(l+1)$$

or,

$$r^2\frac{d^2R}{dr^2} + 2r\frac{dR}{dr} - l(l+1)R = 0$$

The solution is of the form $R = r^s$. To determine the exponent s, we insert this solution back into the above ODE. The end result is

$$s(s+1) = l(l+1) \qquad \Rightarrow \qquad s = l \text{ or } s = -l-1$$

or,

$$R(r) = \begin{cases} r^l \\ r^{-(l+1)} \end{cases}$$

Eq. (16.122) also yields:

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left[l(l+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0$$

One can then carry out the following change of variables $x = \cos\theta$ and $y = \Theta(\theta)$ so that the above equation reduces to:

$$(1 - x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \left[l(l+1) - \frac{m^2}{\sin^2\theta}\right]y = 0$$

This equation is the differential equation for associated Legendre polynomials[47]. We then have

$$y = P_l^m(x) \qquad \text{for } l = 0, 1, 2, 3, \dots \quad \text{and } m = -l, -l+1, \dots 0, \dots l = 1, l$$

and

$$\boxed{P_l^m(x) = \frac{(-1)^m}{2^l\, l!}(1 - x^2)^{m/2}\frac{d^{l+m}}{dx^{l+m}}(x^2 - 1)^l}$$

with $m \geq 0$ and $l \geq 0$. The first few polynomials are

$$P_0^0(\cos\theta) \;=\; 1$$

$$P_1^{-1}(\cos\theta) \;=\; \frac{1}{2}\sin\theta$$
$$P_1^0(\cos\theta) \;=\; \cos\theta$$
$$P_1^{+1}(\cos\theta) \;=\; -\sin\theta$$

$$P_2^{-2}(\cos\theta) \;=\; \frac{1}{8}\sin^2\theta$$
$$P_2^{-1}(\cos\theta) \;=\; \frac{1}{2}\sin\theta\cos\theta$$
$$P_2^0(\cos\theta) \;=\; \frac{1}{2}(3\cos^2\theta - 1)$$
$$P_2^{+1}(\cos\theta) \;=\; -3\sin\theta\cos\theta$$
$$P_2^{+2}(\cos\theta) \;=\; 3\sin^2\theta$$

In our case the differential equation for the associated Legendre polynomials, given above, depends on m^2 and is therefore not sensitive to the sign of m. Consequently, $P_l^m(x)$ and $P_l^{-m}(x)$ must be equivalent solutions and hence proportional to each other, and one can show that

$$P_l^{-m}(\cos\theta) = (-1)^m\frac{(l-m)!}{(l+m)}P_l^m(\cos\theta) \tag{16.123}$$

Combining all the results obtained above, we have found that the general solution to Laplace's equation is of the form

$$T(r, \theta, \phi) = \left\{\begin{array}{c} r^l \\ r^{-(l+1)} \end{array}\right\} P_l^m(\cos\theta)\left\{\begin{array}{c} e^{im\phi} \\ e^{-im\phi} \end{array}\right\}$$

where $l = 0, 1, 2, 3, \dots$ and $m = -l, -l+1, \dots, l-1, l$.

When solving the Laplace's equation in spherical coordinates, it is traditional to introduce the spherical harmonics, $Y_l^m(\theta, \phi)$:

$$Y_l^m(\theta, \phi) = (-1)^m\sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta)e^{im\phi} \qquad \text{for } l = 0, 1, 2, 3, \dots \text{ and } m = -l, -l+1, \dots, l-1, l \tag{16.124}$$

---

[47]https://en.wikipedia.org/wiki/Associated_Legendre_polynomials

The phase factor (-1), introduced originally by Condon and Shortley, is convenient for applications in quantum mechanics. Note that Eq. (16.123) implies that

$$Y_l^{-m}(\theta, \phi) = (-1)^m Y_l^m(\theta, \phi)^*$$

where the star means complex conjugation.

The normalization factor in Eq. (16.124) has been chosen such that the spherical harmonics are normalized to one. In particular, these func- tions are orthonormal and complete. The orthonormality relation is given by:

$$\int Y_l^m(\theta, \phi) Y_{l'}^{m'}(\theta, \phi) d\Omega = \delta_{ll'} \delta_{mm'}$$

where $d\Omega = \sin\theta d\theta d\phi$ is the differential solid angle in spherical coordinates.

It is important to note that there are different normalisations for spherical harmonics. In this document we choose:

$$\boxed{Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi}} \qquad \text{for } l = 0, 1, 2, 3, \dots \text{ and } m = -l, -l+1, \dots, l-1, l$$

(16.125)

which is Eq.(7.8.1) in the Schubert, Turcotte & Olson book [1216]. In this case the $(-1)^m$ is inside the P_m^m. The first few spherical harmonics are shown below in the real representation (i.e.cusing $\cos m\phi$ instead of $e^{im\phi}$) [48] [49]:

$$Y_0^0(\theta, \phi) = \sqrt{\frac{1}{4\pi}}$$

$$Y_1^{-1}(\theta, \phi) = \sqrt{\frac{3}{8\pi}} \cos\phi \sin\theta$$

$$Y_1^0(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos\theta$$

$$Y_1^{+1}(\theta, \phi) = -\sqrt{\frac{3}{8\pi}} \cos\phi \sin\theta$$

$$Y_2^{-2}(\theta, \phi) = \sqrt{\frac{15}{32\pi}} \cos(2\phi) \sin^2\theta$$

$$Y_2^{-1}(\theta, \phi) = \sqrt{\frac{15}{8\pi}} \cos\phi \sin\theta \cos\theta$$

$$Y_2^0(\theta, \phi) = \sqrt{\frac{5}{16\pi}} (3\cos^2\theta - 1)$$

$$Y_2^{+1}(\theta, \phi) = -\sqrt{\frac{15}{8\pi}} \cos\phi \sin\theta \cos\theta$$

$$Y_2^{+2}(\theta, \phi) = \sqrt{\frac{15}{32\pi}} \cos(2\phi) \sin^2\theta$$

replace those my complex ones !

Another normalisation is sometimes used:

$$Y_l^m(\theta, \phi) = \sqrt{(2l+1) \frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi} \qquad \text{for } l = 0, 1, 2, 3, \dots \text{ and } m = -l, -l+1, \dots, l-1, l$$

(16.126)

---

[48]https://en.wikipedia.org/wiki/Table_of_spherical_harmonics
[49]https://mathworld.wolfram.com/SphericalHarmonic.html

with

$$\frac{1}{4\pi} \int Y_l^m(\theta, \phi) Y_{l'}^{m'}(\theta, \phi) d\Omega = \delta_{ll'} \delta_{mm'}$$

Remark. *In [1521] the authors use a normalized associated Legendre polynomial that is related to the associated Legendre polynomial P_l^m as:*

$$p_{lm}(\theta, \phi) = \sqrt{\frac{2l+1}{2\pi(1+\delta_{m0})}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta)$$

Note the absence of the $(-1)^m$ term and the presence of the kronecker delta in the denominator.

📚 Relevant Literature: SHTools: Tools for Working with Spherical Harmonics [1454]

Chapter 17

Geodynamics GEO3-1313 syllabus (Utrecht University)

This course has officially be retired and is no more given by me with this content in the UU Batchelor program (April 2024).

What follows was written by Arie van den Berg and was/is used as the syllabus for the 3rd year geodynamics course at Utrecht University. It is reproduced with Arie's permission and has been slightly modified by me.

17.1 Introduction

The internal constitution of the Earth has been investigated systematically from the nineteenth century on. With the advent of seismological instrumentation for the registration of tele-seismic events, by the end of that century, the main tool for obtaining direct information about distribution of the material properties controling seismic wave propagation became available. Before this, mainly global properties could be determined from gravity and magnetic field observations, astronomical data and indications about the heatflow from the Earth's interior. As a result of the early seismological investigations the main internal structure of the Earth was revealed within the first few decades of the twentieth century with the discovery of the earth's core in 1906 by Oldham and Gutenberg (1912) and the solid inner core in 1936 by Lehmann.

From the radial distribution of the seismic velocity profile, obtained by processing the tables of traveltime versus epicentral distance, Williamson and Adams (1923) [1463] made a first estimate of the density profile for a compressible homogeneous mantle model, consistent with the total mass of the Earth and obtained at the same time strong indication for a high density core, compositionally distinct from the mantle. They concluded that "It is therefore impossible to explain the high density of the Earth on the basis of compression alone. The dense interior cannot consist of ordinary rocks compressed to a small volume; we must therefore fall back on the only reasonable alternative, namely, the presence of a heavier material, presumably some metal, which, to judge from its abundance in the Earth's crust, in meteorites and in the Sun, is probably iron."

Bullen[1] (1975) [182] further refined the analysis and showed the assumption of a homogeneous mantle to be inconsistent with the known moment of inertia of the Earth. In the 1940s and 1950s he introduced a global divison of the Earth in concentric shells, labelled A through G, ranging from the Earth's crust (A), bounded by the moho discontinuity, to the inner core (G). Region C between,

---

[1]Keith Edward Bullen (29 June 1906 - 23 September 1976) was a New Zealand-born mathematician and geophysicist. He is noted for his seismological interpretation of the deep structure of the Earth's mantle and core.

roughly 400km and 900km, characterized by rapid increase of the seismic velocities, was identified by Bullen as a transition region between the upper mantle region B and a homogeneous lower mantle, region D. The deduced inhomogeneity of the mantle was projected by Bullen in this C region. E through G were used to label subdivisions of the core. Region E indicated the liquid, adiabatic outer core, F a transition region between inner and outer core and G the solid inner core. Birch (1952) [96] published improved equations of state, based on finite-strain theory, thereby giving a more firm physical basis to interpretation of available data in terms of a compressible medium.

In the second half of the twentieth century the resolution and accuracy of the models were further improved using continuously improved seismological observations and a growing data set. It also became possible to obtain independent information about the radial density distribution from spectral analysis of radial eigen-vibrations of the Earth after very large earthquakes. This development resulted in the publication of the Preliminary Reference Earth Model (PREM) by Dziewonski and Anderson (1981) [380] which still serves as a global reference.

The improved seismologial models indicated that the continuous rapid velocity increase in the transition zone (C) was actually a succession of several abrupt changes, confirming radial inhomogeneity in mineral phase and possibly in chemical composition of the mantle.

From geological and cosmochemical arguments a probable composition of the Earth had been derived consisting of a mantle with major element composition dominated by magnesium-iron silicates and an iron-nickle core with a small amount of lighter elements mixed in, most likely including mainly sulphur. In the 1960s this resulted in the definition of a so-called pyrolitic composition of the mantle by Ringwood which could explain the main mantle petrological observations regarding the complementary nature of basalts and ultra mafic mantle rocks found in ophiolites, kimberlites and mantle perioditite bodies (Ringwood, 1975 [1144]).

In experimental high-pressure and temperature work on the candidate mantle materials a series of phase transitions were found at pressure and temperature values relevant for the Earth's mantle which could be related to the seismic discontinuities revealed by the seismological data. From these the most prominent at approximately 410 and 660km depth were identified as the phase transition of the olivine component $(Mg, Fe)_2 SiO_4$ of the pyrolitic mantle to a denser wadsleyite crystal structure and, at 660km, a transition (dissociation) from a γ-spinel (known as ringwoodite) structure to a two-phase assemblage, post-spinel, i.e. magnesium-iron perovskite, $(Mg, Fe)SiO_3$ and wüstite $(Mg, Fe)O$.

It was also found that the 660 km boundary corresponds to an endothermic phase transition which would have implications for large scale circulation in the mantle, leading to long-standing speculations about the degree of layering in mantle convection (Christensen & Yuen (1985) [269], Albarede & van der Hilst (2002) [4]), http://www.mantleplumes.org.

A more recent development in this area is the discovery of a new phase transition of magnesium-perovskite to a denser form for pressure temperature conditions, approximately 125 GPa 2500 K, relevant for the D" layer close to the core-mantle boundary (Lay *et al.* , 2005, van der Hilst *et al.* , 2007).
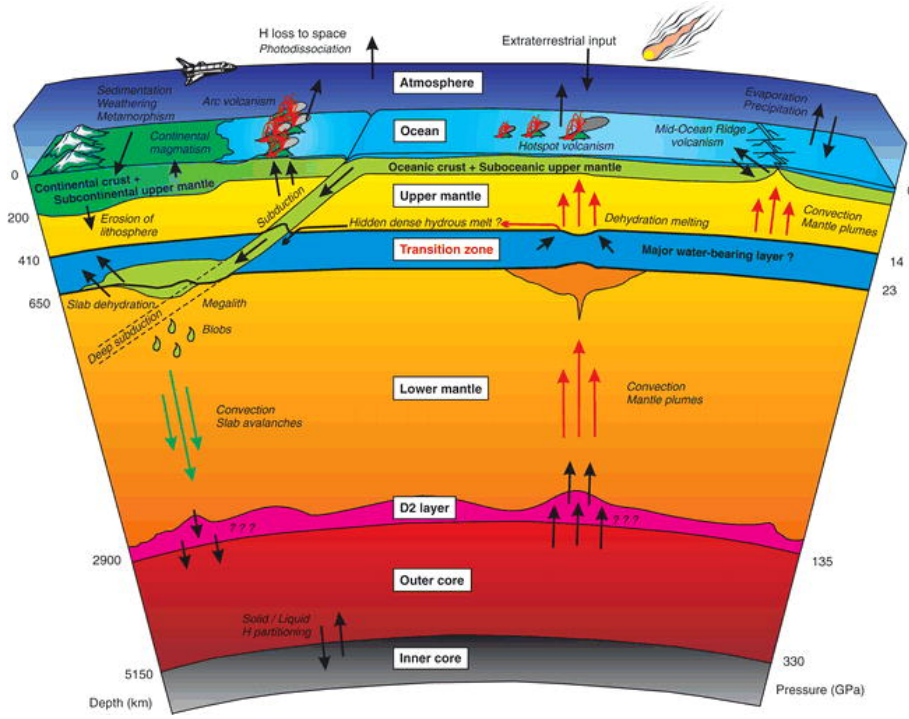
Effect of water on the phase relations in Earth's mantle and deep water cycle,

Litasov & Ohtani (2007) [839]

In the following sections the density distribution in the Earth's interior is treated in relation to the gravity field and internal pressure distribution of a self-gravitating compressible planet model and the link is shown with results from theoretical mineral physics and high pressure-temperature experimental data for mantle materials.

17.2 Global internal structure and temperature of the Earth

To understand the Earth's internal dynamics and evolution we need to know its internal structure and material properties. What do we know about Earth's global internal structure?

For a substance of given chemical composition, the material properties are determined by temperature and pressure. A full understanding of the Earth's internal dynamics therefore requires that we know the internal distribution of composition, temperature and pressure as illustrated in the following figure:
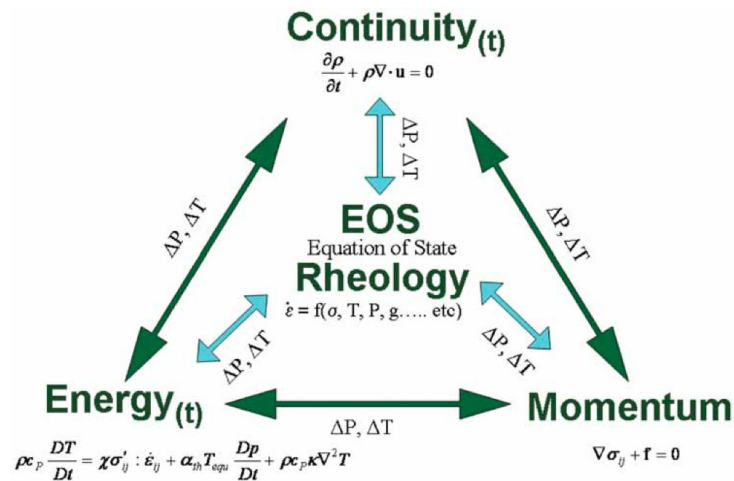
Illustration of the central roles that pressure, temperature and density play in geophysics.

Taken from Regenauer-Lieb *et al.* , Phil. Mag., 2006 [1122].

The internal pressure distribution is directly linked with the Earth's own internal gravity field and density distribution because the local pressure gradient equals the local gravity acceleration times the density (see problem 6). In Section 17.4 density, gravity and pressure are treated together in a consistent way.

If the internal pressure distribution is known we can relate sharp transitions in the physical parameters as shown in the PREM model, illustrated in the following figure to phase transitions, solid-solid or solid-liquid, in the Earth's deep interior.
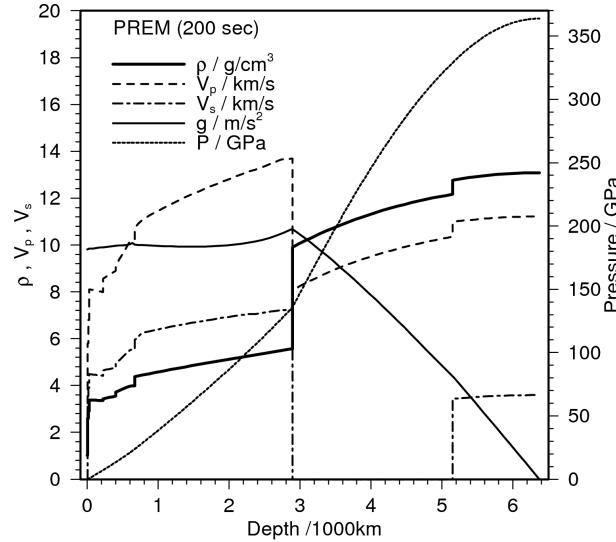
Radial (depth)distribution of density ρ, seismic velocities v_p and v_s, gravity acceleration g and pressure P

in the PREM model (Dziewonski and Anderson (1981) [380]).

Phase transitions in 'candidate' materials for the Earth's interior are investigated under high pressure and temperature conditions in HPT laboratory experiments [2]. Using theoretical mineral physics models the complete mineral phase diagram of mantle silicates can, in principle, be constructed from a limited set of experimental data (Stixrude & Lithgow-Bertelloni (2005) [1298, 1297] Jacobs & de Jong (2007) [662]). To constrain the possible candidate materials we also need to know the internal distribution of the Earth's chemical composition. Such composition models are derived from geological evidence and cosmochemical considerations.

Table I of (Dziewonski and Anderson (1981) [380]) gives an expression for the density as a function of the radius r, which I have turned into a python function:

```python
def prem_density(radius):
    x=radius/6371.e3
    if radius>6371e3:
        densprem=0
    elif radius<=1221.5e3:
        densprem=13.0885-8.8381*x**2
    elif radius<=3480e3:
        densprem=12.5815-1.2638*x-3.6426*x**2-5.5281*x**3
    elif radius<=3630.e3:
        densprem=7.9565-6.4761*x+5.5283*x**2-3.0807*x**3
    elif radius<=5600.e3:
        densprem=7.9565-6.4761*x+5.5283*x**2-3.0807*x**3
    elif radius<=5701.e3:
        densprem=7.9565-6.4761*x+5.5283*x**2-3.0807*x**3
    elif radius<=5771.e3:
        densprem=5.3197-1.4836*x
    elif radius<=5971.e3:
```

---

[2]Deep Earth pressure and temperature conditions can be produced in a Diamond Anvil Cell (DAC), see `http://en.wikipedia.org/wiki/Diamond_anvil_cell`.

```python
            densprem=11.2494−8.0298∗x
        elif radius <=6151.e3:
            densprem=7.1089−3.8045∗x
        elif radius <=6291.e3:
            densprem=2.6910+0.6924∗x
        elif radius <=6346.e3:
            densprem=2.6910+0.6924∗x
        elif radius <=6356.e3:
            densprem=2.9
        elif radius <=6368.e3:
            densprem=2.6
        else:
            densprem=1.020
    return densprem∗1000
```
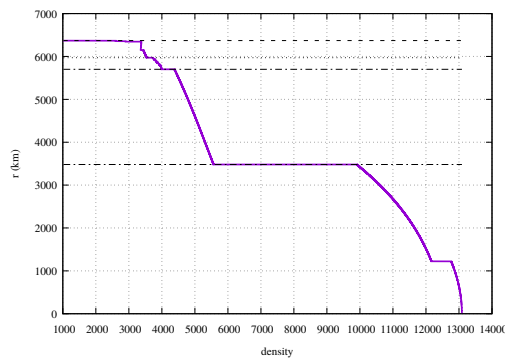

Top: PREM density as computed with the function above. Code available in /images/prem/;
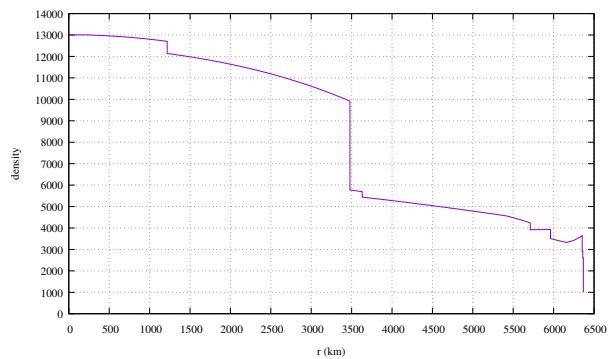
Bottom: ak135 density from [736]. Data available in /images/ak135/

Early models of the Earth's density

The total Earth mass $M_\oplus$ and average density $\langle\rho\rangle$ were not known before independent measurement of Newton's gravitational constant by Cavendish, (see Section 17.4). When the average density had been determined as approximately $5.5\cdot 10^3$ kg $\cdot$ m^{-3} it became clear, from the lower density of surface rocks of around $2.7\cdot 10^3$ kg $\cdot$ m^{-3}, that the Earth's interior must consist of higher density material.

Besides the mass or average density the (average) moment of inertia I (defined in Section 17.3) provides a constraint on the radial distribution of density.

These two integral parameter values have been applied in several two-parameter models for the radial density distribution of the Earth. At the end of the nineteenth century Wiechert[3] assumed that the compressibility of Earth materials would be negligible to first approximation and that Earth's high mean density was due to a dense, probably metallic, core. He assumed an iron core based on astronomical evidence of high iron content of the sun's outer layers (see also Section 17.5.4).

Wiechert considered in particular layered spherically symmetric models consisting of two uniform layers, core and mantle. Since the radius of the Earth's core had not yet been determined by seismology, Wiechert used the core radius R_c and density ρ_c as unknown parameters to be determined from the known data. Wiechert assumed the density of the mantle to be $\rho_m = 3.2\cdot 10^3$kg $\cdot$ m^{-3} and using known values for M and I he derived for the radius of the core $R_c/R = 0.779$ corresponding to a mantle depth of about 1400 km and a core density $\rho_c = 8.2\cdot10^3$ kg $\cdot$ m^{-3}. This model is investigated in problem 3.

Later, after $R_c/R = 0.545$ had been determined using seismic data, Jeffreys substituted the known value of the core radius and derived for the mantle and core densities $\rho_c = 12.6\cdot 10^3$ kg $\cdot$ m^{-3}

---

[3]Emil Johann Wiechert (26 December 1861 - 19 March 1928) was a German physicist and geophysicist who made many contributions to both fields, including presenting the first verifiable model of a layered structure of the Earth and being among the first to discover the electron.

and $\rho_m = 4.14 \cdot 10^3 \text{ kg} \cdot \text{m}^{-3}$ (Bullen (1975) [182]). This model is investigated in problem 4.

The (radially averaged) density distribution in the Earth remains a topic of research [735].

17.3 The moment of inertia of a spherically symmetric density distribution

The moment of inertia I of a point mass of mass m, with respect to a given rotation axis is defined as $I = md^2$ where d is the distance from the point mass to the axis. This quantity relates the angular velocity ω, about the rotation axis, to the angular momentum J, of the point mass, in $J = I\omega$. This is an analogous relation as the one between the linear momentum p and the linear velocity v, $p = mv$. For an extended mass distribution in a volume V, a moment of inertia tensor, I_{ij}, relating the angular momentum vector $\mathbf{J}$ to the rotation vector $\mathbf{\Omega}$ can be defined as $J_i = I_{ij}\Omega_j$, where the summation convention for repeated indices is implied. This tensor is described by a 3×3 matrix defined by volume integration over point masses in the volume. Here we only consider spherically symmetric mass distributions where the moment tensor is isotopic, $I_{ij} = I\delta_{ij}$, with scalar coefficient I. [4] In simple terms, the moment of inertia is the same for any rotation axis through the centre of the spherically symmetric body.

The moment of inertia I can be determined from Earth's global gravity field and the precession rate of the rotation axis determined from astronomical data, see Bullen, *The Earth's density*, 1975. The principal moments of inertia can also be calculated with the hydrostatic equilibrium figure of the Earth [841].

The scalar moment of inertia is defined as a volume integral over point masses,

$$I = \int_V \rho(\vec{r})d(\vec{r})^2 dV. \tag{17.1}$$

where $d(\vec{r})$ is the distance from point $\vec{r}$ to the rotation axis.

For a *spherically symmetric* body of finite volume, it is often expressed in terms of the total mass M, the outer radius R and a prefactor f as,

$$\boxed{I = fMR^2} \tag{17.2}$$

Body	Value	Source
Earth	0.3307	[3]
Mars	0.3662 ± 0.0017	[4]
Mercury	0.346 ± 0.014	[5]
Moon	0.3929 ± 0.0009	[6]
Venus	unknown	

Taken from Wikipedia[5]

We have seen that the planetary mass and surface density were used to constrain models for the interior density distribution. These models are further constrained by the planets moment of inertia

---

[4]δ_{ij} is the Kronecker delta, i.e. $\delta_{ij} = 1$ for $i = j$ and zero otherwise.
[5]https://en.wikipedia.org/wiki/Moment_of_inertia_factor

I that can be determined from (satellite) geodetic and astronomical observations. For Earth the following values for the total mass and moment of inertia prefactor have been found,

$$
\begin{aligned}
M &= 5.97 \cdot 10^{24} \text{kg} \\
I &= 0.3307 M R^2 \text{kg m}^2
\end{aligned}
$$

where $R = 6371$km is the mean radius. The observed moment of inertia prefactor $f = 0.3307$ is smaller than the value 0.4 for a homogeneous sphere (see problem 2), another indication of mass concentration towards the earth's centre.

---

problem: 1. *Derive the following expression for the moment of inertia of a spherically symmetric Earth model with outer radius R,*

$$
I = \frac{8\pi}{3} \int_0^R \rho(r) r^4 dr \tag{17.3}
$$

Hint: use the symmetry and compute $I = \frac{1}{3}(I_x + I_y + I_z)$, where I_x is the moment of inertia with respect to a rotation axis coinciding with the x-axis.

---

problem: 2. *Derive from Eq. (17.3) the value of the prefactor f of the moment of inertia for a uniform sphere. answer: $f = 2/5$.*
In general the moment of inertia prefactor f is an indicator of the degree of mass concentration towards the centre of a spherically symmetric mass distribution. Endmembers of mass concentration are a) a concentrated central point mass and b) all mass concentrated on a spherical surface of zero thickness.
Verify that the moment of inertia of the point mass endmember equals zero and that for the prefactor for a spherical shell of vanishing thickness we have $f = \frac{2}{3}$.

Add bit of theory for delta function is sph coords

Wiechert's two-layer model with a distinct core is constrained by the moment of inertia prefactor f, the mantle radius R and density ρ_m and the total mass M or, equivalently, the mean density $\langle \rho \rangle$. Expressions for the core radius R_c and density ρ_c can be formulated for this model as specified in the following exercise (Bullen, 1975).

for 2025: add figure for pb 2,3. Add plot of rho(r) for pb 3.

problem: 3. *Derive a 2-parameter model for the earth's 1-D radial density distribution $\rho(r)$ consisting of two uniform layers (core and mantle) of radius R_c and R respectively and with contrasting uniform densities ρ_c and ρ_m for core and mantle respectively. Assume ρ_m to be known, leaving ρ_c and R_c as unknown parameters that can be determined from the known moment of inertia prefactor f and the average density $\langle \rho \rangle$.*
Compute the total mass M
Compute the average density and arrive at:

$$\langle \rho \rangle = \frac{3}{R^3} \int_0^R \rho(r) r^2 dr \tag{17.4}$$

Use $I = fMR^2$ and the total mass to arrive at:

$$fR^5 \langle \rho \rangle = 2 \int_0^R \rho(r) r^4 dr \tag{17.5}$$

Derive the following expressions for R_c and ρ_c,

$$\frac{R_c}{R} = \left(\frac{\frac{5}{2} f \frac{\langle \rho \rangle}{\rho_m} - 1}{\frac{\langle \rho \rangle}{\rho_m} - 1} \right)^{1/2} , \quad \rho_c = \rho_m \left\{ 1 + \left(\frac{R}{R_c} \right)^3 \left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right) \right\} \tag{17.6}$$

In Bullen's two-layer model the core radius is assumed to be known from seismology. For this model the mantle and core densities can be expressed in the known parameters in the following problem.

problem: 4. *Assume the core radius R_c to be a known parameter in the following. Derive a 2-parameter model for the earth's 1-D radial density distribution $\rho(r)$ consisting of two uniform layers (core and mantle), with a core and mantle radius R_c and R and different uniform densities ρ_m and ρ_c for mantle and core. Express the parameters ρ_m and ρ_c in terms of the mass and moment of inertia.*
Hint: compute M first, then I, as a function of all other parameters. Establish a relationship of the form $(M, I)^T = A \cdot (\rho_c, \rho_m)^T$ where A is a 2×2 matrix.
Solution: in matrix-vector format,

$$\begin{pmatrix} \rho_c \\ \rho_m \end{pmatrix} = \frac{4\pi}{3\Delta} \begin{pmatrix} \frac{2}{5}(R^5 - R_c^5) & -(R^3 - R_c^3) \\ -\frac{2}{5}R_c^5 & R_c^3 \end{pmatrix} \begin{pmatrix} M \\ I \end{pmatrix} \tag{17.7}$$

where the determinant $\Delta = \frac{32\pi^2}{45} \left(R_c^3(R^5 - R_c^5) - R_c^5(R^3 - R_c^3) \right)$.

Carry out live demo of python code. Code is in images/geodynamics

problem: 5. *SKIP THIS PROBLEM. The numerical value of the interim expressions in (17.7) exceeds the magnitude of single precision real type variables in computer programs, that are limitid to approximately $1.7 \cdot 10^{38}$. A work around for this problem may be to use double precision real variables that have a higher maximum magnitude of about 10^{308}.*

An alternative solution is to switch to using non-dimensional parameters, denoted by primes, in the following way: define $R_c' = R_c/R$, $M_0 = 4/3 \cdot \pi R^3 \rho_0$ and $M = M_0 \cdot M/M_0 = M_0 \cdot M'$, $\rho_c = \rho_0 \rho_c'$, $\rho_m = \rho_0 \rho_m'$ and express the moment of inertia in the reference density ρ_0 and outer radius as, $I = fMR^2 = f4/3 \cdot \pi R^5 \rho_0$. With these definitions rewrite (17.7) into the non-dimensional form,

$$\begin{pmatrix} \rho_c' \\ \rho_m' \end{pmatrix} = \frac{16\pi^2}{9\Delta'} \begin{pmatrix} \frac{2}{5}(1 - R_c'^5) & -(1 - R_c'^3) \\ -\frac{2}{5}R_c'^5 & R_c'^3 \end{pmatrix} \begin{pmatrix} M' \\ f \end{pmatrix} \tag{17.8}$$

where the determinant $\Delta' = \frac{32\pi^2}{45}\left(R_c'^3(1 - R_c'^5) - R_c'^5(1 - R_c'^3) \right)$.

17.4 Density, gravity and pressure in the Earth

In the Earth's mantle major solid state phase transitions occur in the silicate material which constitutes the planetary mantle outside the metallic iron/nickle core. These phase transitions are induced by the increase in the static pressure from a 1 bar (10^5Pa) atmospheric value at the Earth's surface to $136 \cdot 10^9$Pa at the core mantle boundary at a depth of approximately 2900km. Phase transitions in the Earth's interior are associated with changes in the elastic wave velocities that can be deduced from seismological observations. In high pressure experiments, phase transitions in candidate mantle silicates can be studied and correlated with the seismological data to constrain the mineralogy and pressure/temperature distribution in the mantle. Knowledge of the internal material constitution of the Earth, such as the mineral phase, is a requirement for understanding the main geodynamical processes that determine Earth's evolution.

Density and pressure inside the Earth are linked with self-gravitation. This means that the hydrostatic or lithostatic pressure is a direct result of the gravity field generated by the Earth's own mass distribution. The lithostatic pressure can be expressed as the weight of a column of unit cross-sectional area extending from zero depth, at the Earth's surface, to the depth z of the evaluation point,

$$P(z) = \int_0^z \rho(z')g(z')dz' \tag{17.9}$$

where ρ is the mass density and g is the magnitude of the gravitational acceleration.

The gravity field defining g is generated by the Earth's own density distribution. Weak periodic gravity 'perturbations' are generated by celestial bodies, expressed in the external tides, both ocean tides and solid earth tides. The main tides are generated by the Earth's moon and by the Sun.

In the following section expressions for the gravity field in terms of the density distribution are given, based on Newton's law of gravitation.

In the description of the density distribution we will first neglect the role of self-compression and consider a number of one-dimensional (1-D), spherically symmetric, parameterized density distributions. Self-compression and compressibility are then treated in section 17.5.2. Self-compression and finite compressibility result in a continuous increase of density with pressure in agreement with several geophysical observations.

---

problem: 6. *Derive the expression (17.9) (where the depth z is not to be confused with a cartesian coordinate) for the lithostatic pressure in a spherically symmetric planet from the elastostatic equation for a static medium,*

$$\partial_j \sigma_{ij} + \rho g_i = 0 \qquad \Leftrightarrow \qquad \vec{\nabla} \cdot \boldsymbol{\sigma} + \rho \vec{g} = \vec{0} \tag{17.10}$$

Hint: Assume hydrostatic conditions where the stress tensor can be written as $\sigma_{ij} = -P\delta_{ij}$, with δ_{ij} the Kronecker delta, and derive from equation (17.10) for the pressure gradient, $\vec{\nabla}P = \rho\vec{g}$.

Gravity field of a mass distribution

Newton formulated the attraction force acting on a point mass m_0, located in a point with position vector $\vec{r} = (x, y, z)$, with x, y, z the cartesian coordinates, from a second point mass m_1 located at $\vec{r}_1 = (x_1, y_1, z_1)$, illustrated in the following figure as,

$$\vec{F}(\vec{r}) = \frac{\mathcal{G}m_0 m_1}{|\vec{r}_1 - \vec{r}|^2}\, \vec{e}_{\vec{r}\vec{r}_1} \tag{17.11}$$
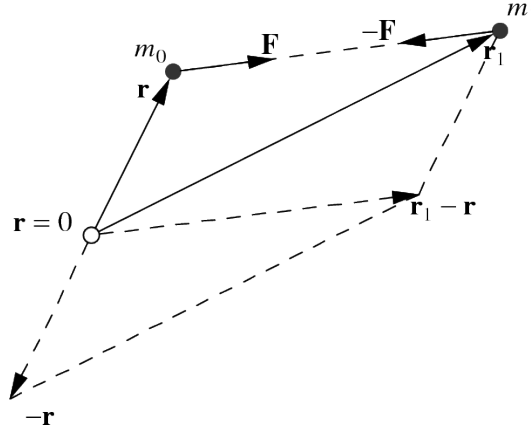
Vector diagram of the gravitational forces acting on the two point masses m_0, m_1 in vector locations $\vec{r}$ and $\vec{r}_1$ respectively. From the expression for the gravity field (17.11) it follows that the forces on both masses are of equal magnitude and in opposite direction.

Where $\vec{e}_{\vec{r}\vec{r}_1}$ is the unit vector in $\vec{r}$ pointing towards $\vec{r}_1$ and $\vec{F}(\vec{r}_1) = -\vec{F}(\vec{r})$. $\vec{r}, \vec{r}_1$ are the position vectors of the two point masses and $|\vec{r}_1 - \vec{r}| = \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}$ is the distance between the points $\vec{r}$ and $\vec{r}_1$. $\mathcal{G}$ is the gravitational constant $\mathcal{G} \simeq 6.67 \times 10^{-11} \mathrm{N\,m^2\,kg^{-2}}$, m_0, m_1 the mass of the respective pointmasses.

This gravitation effect is usually specified as a gravitation force per unit mass or acceleration vector $\vec{g}$,

$$\vec{g}(\vec{r}) = \frac{\mathcal{G}m_1}{|\vec{r}_1 - \vec{r}|^2} \ \vec{e}_{\vec{r}\vec{r}_1} \tag{17.12}$$

It can be verified by inspection that the acceleration vector field can be written as the gradient of a scalar potential field $U(\vec{r})$ (i.e. the potential energy per unit mass) with

$$\vec{g} = -\vec{\nabla}U = \left(-\frac{\partial U}{\partial x}, -\frac{\partial U}{\partial y}, -\frac{\partial U}{\partial z}\right),$$

in Cartesian coordinates (see problem 9), and

$$U(\vec{r}) = -\frac{\mathcal{G}m_1}{|\vec{r}_1 - \vec{r}|} \tag{17.13}$$

The gravity acceleration and corresponding potential field are additive such that the total force or potential of a collection of N point masses is obtained by summation over individual point contributions,

$$\vec{g}(\vec{r}) = \sum_j^N \frac{\mathcal{G}m_j}{|\vec{r}_j - \vec{r}|^2} \ \vec{e}_{\vec{r}\vec{r}_j}, \qquad U(\vec{r}) = -\sum_j^N \frac{\mathcal{G}m_j}{|\vec{r}_j - \vec{r}|} \tag{17.14}$$

With this definition and sign convention the potential field of a point source in the origin is represented by a potential well ($U(\vec{r}) < 0$). This is known as Coulomb's law and the equivalent form for a continuous mass distribution of density ρ (mass per unit volume) contained in a volume V is,

$$\vec{g}(\vec{r}) = \int_V \frac{\mathcal{G}\rho(\vec{r}')}{|\vec{r}' - \vec{r}|^2} \ \vec{e}_{\vec{r}\vec{r}'} \ dV(\vec{r}'), \qquad U(\vec{r}) = -\int_V \frac{\mathcal{G}\rho(\vec{r}')}{|\vec{r}' - \vec{r}|} \ dV(\vec{r}') \tag{17.15}$$

Besides the integral expression for the gravity field defined in (17.15) there is also the differential form using the second order partial differential equations of Laplace and Poisson. It can be shown by verification (see hereafter) that U in (17.15) satisfies Poisson's equation,

$$\vec{\nabla}^2 U = 4\pi\mathcal{G}\rho \qquad (17.16)$$

which reduces to Laplace's equation $\vec{\nabla}^2 U = 0$ outside the mass distribution in V (where $\rho = 0$).

To show that U in (17.15) satisfies Poisson's equation integrate the normal component of the acceleration field over an arbitrary closed surface S enclosing V and change the order of integration for the volume and surface integral.

$$\int_S \vec{\nabla} U(\vec{r}) \cdot \vec{n}\, dA(\vec{r}) = -\int_V \mathcal{G}\rho(\vec{r'}) \left\{ \int_S \vec{\nabla} \left(\frac{1}{|\vec{r'} - \vec{r}|} \right) \cdot \vec{n}\, dA(\vec{r}) \right\} dV(\vec{r'}) \qquad (17.17)$$

The surface integral on the right is independent of the choice of the surface S as long as it contains $\vec{r'}$. We therefore replace this surface by a sphere of radius R centered at r' and find for the surface integral the value -4π.

Next we apply the Gauss divergence theorem to the left hand surface integral to obtain,

$$\int_V \vec{\nabla}^2 U\, dV = \int_V 4\pi\mathcal{G}\rho\, dV \qquad (17.18)$$

Note that the surface has been contracted on the volume V to obtain (17.18). Since the surface and enclosed volume are arbitrary we obtain the Poisson equation,

$$\vec{\nabla}^2 U = 4\pi\mathcal{G}\rho \qquad (17.19)$$

In Newton's time the numerical value of $\mathcal{G}$ had not been determined yet. As a result it was not possible to determine the mass of the Earth $M_\oplus$ by measuring the gravitation force of the Earth on a known 'test mass'. This way only the value of $GM_\oplus$ could be determined. Only with the experiment named after Cavendish (1798) [6] it became possible to measure $\mathcal{G}$ directly, in a torsion balance experiment, by determining the gravitational attraction of two closely spaced test masses shown here:

Fig. 1
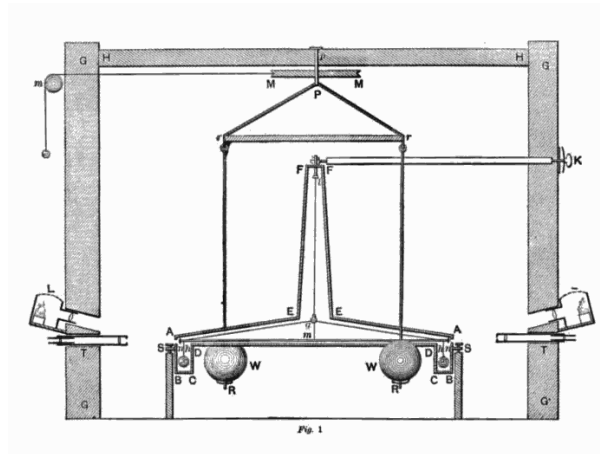
Since Cavendish many experiments have been conducted in order to determine the value of this constant:

---

[6] http://en.wikipedia.org/wiki/Cavendish_experiment
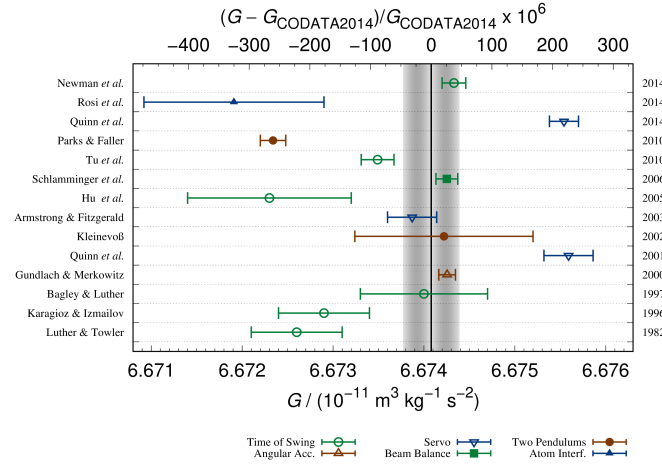
The recommended value is $\mathcal{G} = 6.67430(15) \cdot 10^{-11} \mathrm{m}^3 \mathrm{kg}^{-1} \mathrm{s}^{-2}$, see for instance https://physics.nist.gov/cgi-bin/cuu/Value?bg.

Multipole expansion

The idea behind the multipole expansion is simple: the denominator in Eq. (17.13) can be rewritten in such a way that it can be expanded as a Taylor series. We have

$$\frac{1}{|\vec{r} - \vec{r}\,'|} = \frac{1}{\sqrt{(\vec{r} - \vec{r}\,') \cdot (\vec{r} - \vec{r}\,')}} = \frac{1}{\sqrt{\vec{r} \cdot \vec{r} - 2\vec{r} \cdot \vec{r}\,' + \vec{r}\,' \cdot \vec{r}\,'}} = \frac{1}{r^2} \frac{1}{\sqrt{1 - 2\frac{\vec{r} \cdot \vec{r}\,'}{r^2} + \left(\frac{r'}{r}\right)^2}}$$

The potential can be expanded in a series of Legendre polynomials.

$$(1 - 2XZ + Z^2)^{-1/2} = \sum_{n=0}^{\infty} Z^n P_n(X)$$

valid for $|X| \le 1$ and $|Z| \le q$. The coefficients P_n are the Legendre polynomials of degree n.

FINISH!

Gravitational potential energy

For two pairwise interacting point particles, the gravitational potential energy $\mathcal{U}$ is given by

$$\mathcal{U} = \frac{-\mathcal{G}Mm}{R}$$

where M and m are the masses of the two particles, R is the distance between them. Close to the Earth's surface, the gravitational field is approximately constant, and the gravitational potential energy of an object reduces to

$$\mathcal{U} = mgh$$

where m is the object's mass, $g = \mathcal{G}M_E/R_E^2$ is the gravity of Earth, and h is the height of the object's center of mass above a chosen reference level.

Let us talk units

The SI units for (gravity) acceleration are $\mathrm{m\,s^{-2}}$. However in the context of gravity, we will rarely encounter these.

The Gal is the commonly used unit in gravimetry:

$$0.01\mathrm{m\,s^{-2}} = 1\mathrm{Gal}$$

and often measurements are given in mGal or μGal.

As such, the acceleration due to Earth's gravity at its surface is 976 to 983 Gal, the variation being due mainly to differences in latitude and elevation.

Gravity Force Inside a Spherical Shell

In classical mechanics, the *shell theorem* gives gravitational simplifications that can be applied to objects inside or outside a *spherically symmetrical* body.

Isaac Newton proved the shell theorem and stated that:

1. A spherically symmetric body affects external objects gravitationally as though all of its mass were concentrated at a point at its centre.

2. If the body is a spherically symmetric shell (i.e., a hollow ball), no net gravitational force is exerted by the shell on any object inside, regardless of the object's location within the shell.
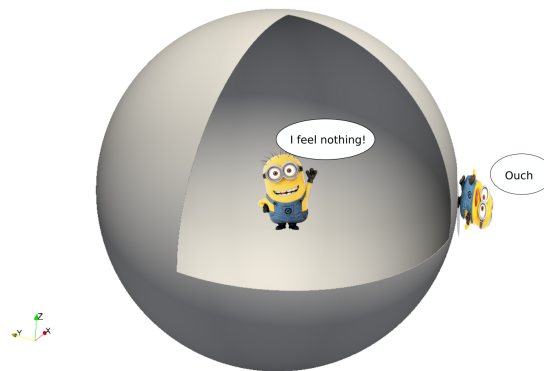
These two propositions are not easy to prove. The second one is very important: it states that if I stand mid-mantle at a radius of, say, 5000km, the 1371km-thick shell of rock above me does not contribute to the force of gravity that I am feeling. Only the rocks below my feet contribute to this force. At this location we can write

$$\frac{\mathcal{G}mM(r)}{r^2} = ma$$

where $M(r)$ is the mass inside a sphere of radius r. The mass m of my body cancels out, and we obtain

$$\frac{\mathcal{G}M(r)}{r^2} = a$$

The acceleration in this context is often called g and it clearly depends on r so that if density is constant, $M(r) = \frac{4\pi}{3}r^3\rho_0$ and then

$$g(r) = \mathcal{G}\frac{4\pi}{3}r\rho_0$$

It also follows that the gravity acceleration in the center of the planet ($r = 0$) must be zero and the gravity acceleration increases linearly with the distance to the center. If now the density is not constant (but radially symmetric, i.e. $\rho = \rho(r)$) then

$$g(r) = \mathcal{G}\frac{4\pi}{r^2}\int_0^r \rho(r')r'^2 dr'$$

Remember that this is true because of the spherical symmetry!

Nonuniqueness

Given a body with mass M at a distance r from me, the gravitational acceleration that I feel is

$$a = \frac{\mathcal{G}M}{r^2}$$

If the mass is now twice as far (distance $2r$) then

$$a' = \frac{\mathcal{G}M}{(2r)^2} = \frac{1}{4}\frac{\mathcal{G}M}{r^2} = \frac{1}{4}a$$

Because of the inverse square of the distance the acceleration is four times as small.

However, if I now 'make' the mass of the body four times as large and twice as far,

$$a'' = \frac{\mathcal{G}(4M)}{(2r)^2} = \frac{\mathcal{G}M}{r^2} = a$$

There lies a very important fact: There is an inescapabale trade-off between distance and mass.

If gravity is measured at a single point in space nothing certain can be said about what lies below: the object generating the gravity anomaly could be 'close' and not so massive, or 'far' and really massive, both situations potentially leading to the same measurement.
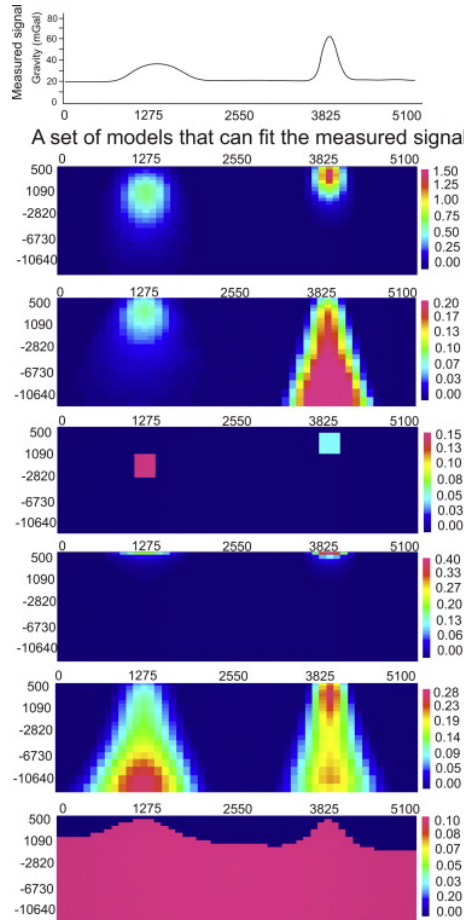
Taken from Meijde, Pail, Bingham, and Floberghagen [910] (2015).

The Netherlands

As explained in Crombaghs, Min, and Hees [306] (2002), in The Netherlands gravity values increase from south to north with about 1 milligal per kilometer. Smallest values occur in Limburg (981,100 milligal), while the largest values occur in Groningen (981,350 milligal). Local variations are limited to 1 milligal over some kilometers.
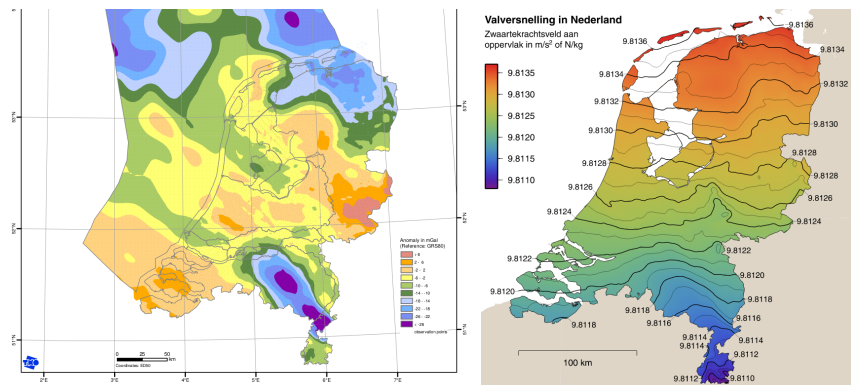
Left: Taken from https://www.nlog.nl/en/gravity-and-magnetic-field. The size of the Bouguer anomaly at a particular location is a measure of the mass deficit or mass excess in the underlying rocks. A mass deficit exists where the stratigraphic succession is composed of relatively light rocks; this yields a negative Bouguer anomaly. A mass excess exists where the stratigraphic succession is composed of relatively heavy rocks, this yields a positive anomaly. Right: Taken from https://upload.wikimedia.org/wikipedia/commons/c/ce/Valversnelling_in_Nederland.svg.
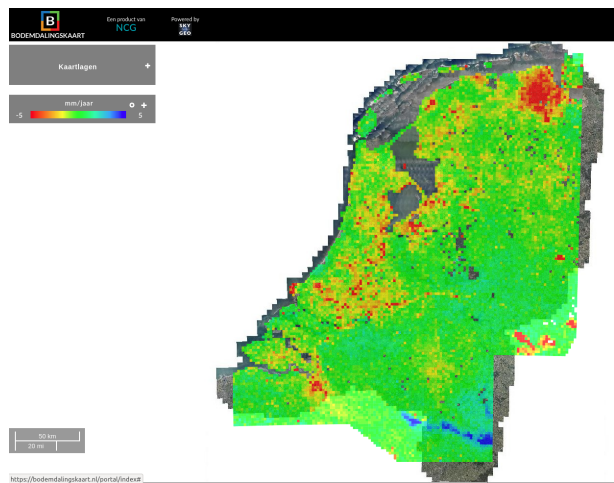
Taken from https://bodemdalingskaart.nl/portal/index. Is the continuous sinking of certain parts of the Netherlands visible in the satellite gravity rate measurements ?

write about Anomalies, Additivity, Inversion

A few problems more to solve

problem: 7. *Verify that the familiar surface value of the Earth's gravity acceleration $g_0 = 9.8$ m/s^2 corresponds to the value of a point mass at the Earth's centre with the same mass as the Earth (see Table).*

	Radius km	Mass kg	Density kg/m^3
Earth	6371	$5.97 \cdot 10^{24}$	5.515×10^3
Moon	1738	$7.34 \cdot 10^{22}$	3.34×10^3
Mars	3394	$6.42 \cdot 10^{23}$	3.93×10^3
Jupiter	71492	$1.9 \cdot 10^{27}$	1.326×10^3
Sun	$6.96 \cdot 10^5$	$1.99 \cdot 10^{30}$	-

Radius-mass parameters of Earth moon and planets.
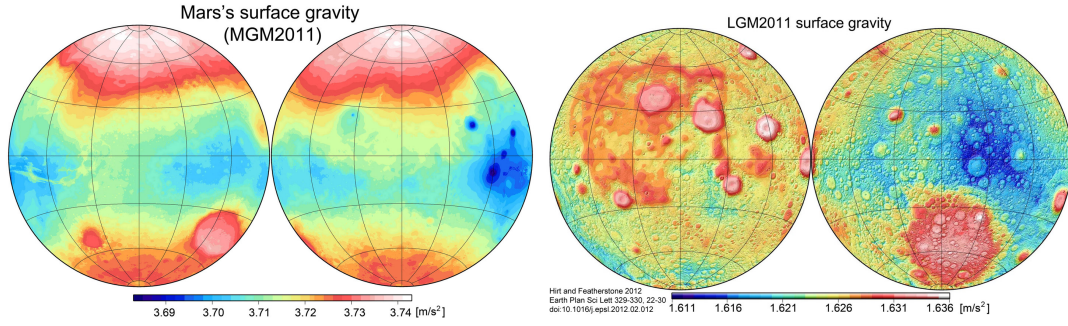
Left: Mars gravity, taken from Hirt *et al.* (2011) [609]; Right: Moon gravity, taken from `https://en.wikipedia.org/wiki/Gravitation_of_the_Moon`

---

problem: 8. *The PREM profile suggests that the magnitude of the gravity aceleration is approximately constant throughout the Earth's mantle. Assume an approximate uniform value of g in the Earth's mantle, equal to the surface value $g_0 \sim 9.8 m/s^2$ and use an approximate average mantle density $\rho_m \sim 4.5 \times 10^3 kg/m^3$ to obtain from Eq. (17.9) an approximation of the static pressure at the core mantle boundary at a depth of 2891km.*

---

problem: 9. *Verify the consistency of the expression for the gravity acceleration and potential of a point mass in (17.12) and (17.13), i.e. prove from these expressions by explicit calculation of the gradient vector from the scalar potential field that $\vec{g} = -\vec{\nabla} U$.*

Hint: specify the potential in (17.13) in Cartesian coordinates (i.e. write explicitely $|\vec{r}_1 - \vec{r}|$) and differentiate the result with respect to the coordinates x, y, z. What is the derivative of $(f(x))^\alpha$ with respect to x? After a few steps you should then arrive at (17.12).

---

problem: 10. *Apply the Poisson equation (17.16) to obtain the gravity field of a point-mass distribution with mass M, described by a Dirac delta function, $\rho(\vec{r}) = M\delta(\vec{r} - \vec{r}_0)$. Where the following property holds for the delta function,*

$$\int_V \delta(\vec{r} - \vec{r}_0)dV = \left\{ \begin{array}{l} 1, \ \vec{r}_0 \in V \\ 0, \ \vec{r}_0 \ni V \end{array} \right. \quad or, \ more \ general \ \int_V f(\vec{r})\delta(\vec{r} - \vec{r}_0)dV = \left\{ \begin{array}{l} f(\vec{r}_0), \ \vec{r}_0 \in V \\ 0, \ \vec{r}_0 \ni V \end{array} \right.$$
$$(17.20)$$

Hint: integrate (17.16) over a spherical volume, centered at $\vec{r}_0$ and apply the Gauss divergence theoreme: for a vector field $\vec{A} = (A_1, A_2, A_3)$ with divergence $\vec{\nabla} \cdot \vec{A} = \frac{\partial A_1}{\partial x} + \frac{\partial A_2}{\partial y} + \frac{\partial A_3}{\partial z}$

$$\int_V \vec{\nabla} \cdot \vec{A} \, dV = \int_{\partial V} \vec{A} \cdot \vec{n} \, dS \qquad (17.21)$$

where ∂V is the closed boundary surface of V.

problem: 11. *Check the dimensional units in (17.16) and verify that the gravitational potential has the dimension of energy per unit mass. This is in agreement with the identification of the gravity potential with the potential (gravitational) energy of a unit mass in the gravity field.* [a]

---

[a]The local potential field value $U(\mathbf{r}_1)$ equals the negative of the (gravitational) potential energy $W(\mathbf{r}_1)$ of a unit point mass positioned at $\mathbf{r}_1$. It can be shown that the change in potential energy ΔW that results from moving a unit mass from $\mathbf{r}_1$ to $\mathbf{r}_2$ follows directly from the potential field values $U(\mathbf{r}_1)$, $U(\mathbf{r}_2)$ and is independent of the path taken between $\mathbf{r}_1$ and $\mathbf{r}_2$. This property defines a so called conservative field U.

To derive this result we compute the potential energy difference as the path (line) integral of the work done by the gravity force field on a unit mass and apply the gradient property $\vec{g} = -\vec{\nabla}U$. The work done by moving a unit point mass from a location $\mathbf{r}_1$ to $\mathbf{r}_2$ is defined by the line integral,

$$\Delta W \quad = \quad \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} = \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{g} \cdot d\mathbf{r} = \int_{\mathbf{r}_1}^{\mathbf{r}_2} -\vec{\nabla}U \cdot d\mathbf{r} = \int_{U(\mathbf{r}_1)}^{U(\mathbf{r}_2)} -dU = -\left(U(\mathbf{r}_2) - U(\mathbf{r}_1)\right) = -\Delta U \quad (17.22)$$

Here the following gradient property has been used, relating the gradient vector to the differential of the scalar potential field,

$$dU = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy + \frac{\partial U}{\partial z}dz = \vec{\nabla}U \cdot d\vec{r} \quad (17.23)$$

The gravitational potential field can thus be defined in terms of the work done by the gravity field to move a unit mass from infinity to the evaluation point.

$$W(\mathbf{r}_1) = \int_{\mathbf{r}_\infty}^{\mathbf{r}_1} \mathbf{g} \cdot d\mathbf{r} = \int_{\mathbf{r}_\infty}^{\mathbf{r}_1} -\vec{\nabla}U \cdot d\mathbf{r} = \int_{U(\mathbf{r}_\infty)}^{U(\mathbf{r}_1)} -dU = -U(\mathbf{r}_1) + U(\mathbf{r}_\infty) = -U(\mathbf{r}_1) \quad (17.24)$$

Where $U(\mathbf{r}_\infty) = 0$ has been used.

The above can be applied in the determination of the escape velocity from the surface of a planet. This is the minimum launch velocity to escape from the planet's gravity field. For a spherically symmetric planet the external gravity potential is given by (17.33). Moving an object from the surface, the gravity potential changes by $\Delta U = U(r) - U(R) = \mathcal{G}M(-\frac{1}{r} + \frac{1}{R})$. Applying an energy conservation argument we require the change in total (potential plus kinetic) energy per unit mass to be: $\Delta E = \Delta U + \Delta K = 0$. With $\Delta K = -v_{ex}^2/2$ we get $v_{esc} = \sqrt{2\mathcal{G}M/R}$.

problem: 12. *Compute the surface escape velocities for different celestial bodies using the parameters given in the Table above.*

To watch:

- Tossing Satellites into Orbit with SpinLaunch: Is that Possible? (3min)
  https://www.youtube.com/watch?v=-BCeanUiKwM

- Can We Throw Satellites to Space? (42min)
  https://www.youtube.com/watch?v=yrc632oilWo

> **problem: 13.** *The potential energy of a self-gravitating planet in its own gravity field is defined in terms of the volume density ρU as,*
>
> $$E = -\int_V \rho U \, dV \qquad (17.25)$$
>
> *Derive the following expression for the potential energy of a spherically symmetric, uniform density model, using the expression for the internal gravity potential defined in (17.32)*
>
> $$E = \frac{8\pi}{5} \mathcal{G} \rho_0 M R^2 \qquad (17.26)$$
>
> *Compute the potential energy value E, assuming a density $\rho_0 = 5.5 \cdot 10^3 \mathrm{kg/m^3}$ and planetary radius $R = 6371 \mathrm{km}$.*
>
> *answer:* $4.4 \cdot 10^{32} \mathrm{J}$

The gravitational energy considered above plays an important role in major compositional differentiation processes that occurred in the early Earth and are still occuring today.

- A so called 'core catastrophe' occured when the iron/nickel core of the Earth differentiated from the silicate mantle in the first few million years after the formation of the Earth in the early solar system. This event has probably freed enough potential energy to melt the mantle completely, resulting in a global magma ocean [7].

- Crystallization of the solid inner core from the liquid outer core, as a result of core cooling, is accompanied by compositional differentiation. The liquid outer core contains a lighter fraction, possibly sulfur, which stays behind in the liquid during freezing of the inner core. The enriched residual liquid near the inner core boundary is less dense than the average liquid of the outer core and this results in a gravitationally unstable layering that induces 'chemically driven' convective flow in the outer core. The potential energy released in this chemical convection is probably an important energy source in powering the geodynamo that generates the Earth's present day magnetic field.

17.5   The gravitational potential for spherical problems

Starting from the Poisson equation,

$$\Delta U = 4\pi \mathcal{G} \rho$$

and using Gauss' theorem (noting that $\Delta U = \vec{\nabla} \cdot \vec{\nabla} U$):

$$\int_V \Delta U \, dV = \int_V \vec{\nabla} \cdot \vec{\nabla} U \, dV = \int_\Gamma \vec{\nabla} U \cdot \vec{n} \, dS = \int_V 4\pi \mathcal{G} \rho \, dV = 4\pi \mathcal{G} \int_V \rho \, dV$$

where $\vec{n}$ is the outward pointing normal vector.

A uniform sphere of mass M and radius a (and therefore density $\rho = M/(4\pi a^3/3)$) has the potential

$$U(r) = \begin{cases} -2\pi \mathcal{G} \rho (a^2 - r^2/3) & r \le a \\ -\mathcal{G}M/r & r \ge a \end{cases} \qquad (17.27)$$

Outside the sphere the potential is Keplerian, while inside it has the form of a parabola; both the potential and its derivative are continuous at the surface of the sphere.

A sphere with density profile

$$\rho(r) = \rho_0 (r/r_0)^{-2}$$

---

[7]https://en.wikipedia.org/wiki/Iron_catastrophe

has the potential

$$U(r) = 4\pi \mathcal{G} \rho_0 r_0^2 \ln(r/r_0) \tag{17.28}$$

---

problem: 14. *Verify* (17.27) *and* (17.28). *Sketch the density field and the resulting gravity field and potential.*

---

problem: 15. *Assume a spherically symmetric non-rotating Earth in hydrostatic equilibrium. In spherical coordinates the divergence of a vector field $\vec{a}(r)$, which only depends on the radius r is*

$$\vec{\nabla} \cdot \vec{a} = \frac{1}{r^2} \frac{d}{dr}(r^2 a_r)$$

a) Prove that the acceleration of gravity at radius r only depends on the mass contained in the sphere of radius R_i. Hint: start from $\vec{\nabla} \cdot \vec{g}$.
b) Assume that the mass of the Earth's core is M_c. Assume a linear density profile for the crust and mantle and determine the acceleration of gravity as a function of the radius in the mantle.

As it turns out, pairs of functions related by Poisson's equation provide convenient building-blocks for galaxy models. Three such functions often used in the literature are listed here; all describe models characterized by a total mass M and a length scale a:

- Plummer (1905) [345]

$$\rho(r) = \frac{3M}{4\pi a^3} \left(1 + \frac{r^2}{a^2}\right)^{-5/2} \qquad U(r) = -\frac{\mathcal{G}M}{\sqrt{r^2 + a^2}}$$

- Hernquist (1990)

$$\rho(r) = \frac{M}{2\pi} \frac{a}{r(r+a)^3} \qquad U(r) = -\frac{\mathcal{G}M}{r+a}$$

- Jaffe (1983)

$$\rho(r) = \frac{M}{4\pi} \frac{a}{r^2(r+a)^2} \qquad U(r) = \frac{\mathcal{G}M}{a} \ln \frac{a}{r+a}$$

(VERIFY?)

17.5.1 The gravity and pressure field for parameterized density models with self-gravitation

In the following problems a number of simple density distributions are investigated that will serve as a reference for models more constrained by geophysical observations to be introduced in later sections. The gravity field can be determined by solving the governing Poisson equation (17.16) using suitable boundary conditions. For the special case of spherically symmetric mass distributions simple 1-D integral expressions can be used to derive the corresponding radial pressure distribution.

problem: 16. *The internal and external gravity field for a simple model of a planet can by derived by solving the Poisson equation (17.16), and applying appropriate boundary conditions to the general solution. Consider a spherically symmetric planet of radius R and uniform density ρ_0.*

1. *Derive expressions for the gravity potential field U and the gravity force field $g = |\mathbf{g}|$ inside and outside the planet.*

   Hints: Solve Poisson's equation in spherical coordinates for the interior $(r \leq R)$ and exterior domain $r \geq R$ separately. The separate solutions for the interior U_{int}, g_{int} and exterior U_{ext}, g_{ext} domain each contain two integration constants which can be determined by applying the following boundary conditions,

$$\lim_{r \to \infty} U_{ext}(r) = 0, \quad \lim_{r \to 0} g_{int}(r) < \infty \qquad (17.29)$$

   Continuity of the gravity accelaration g at the surface $r = R$,

$$g_{int}(R) = g_{ext}(R) \qquad (17.30)$$

   Continuity of the gravity potential U at the surface $r = R$,

$$U_{int}(R) = U_{ext}(R) \qquad (17.31)$$

   Answers

$$g_{int} = \frac{4\pi}{3} \mathcal{G} \rho_0 r \;, \quad U_{int} = \frac{2\pi}{3} \mathcal{G} \rho_0 r^2 - \frac{3}{2} \frac{\mathcal{G} M}{R} \qquad (17.32)$$

   where $M = \frac{4\pi}{3} R^3 \rho_0$ is the planet mass and $\mathcal{G}$ is the gravitational constant.

$$g_{ext} = \frac{\mathcal{G} M}{r^2} \;, \quad U_{ext} = -\frac{\mathcal{G} M}{r} \qquad (17.33)$$

2. *Verify that the external gravity force field is identical to the field of a concentrated point mass at $r = 0$.*

3. *Derive an expression for the radial distribution of the pressure in the planetary interior and compute the central pressure for a case with $\rho_0 = 5.5 \cdot 10^3 \text{kgm}^{-3}$ and $R = 6.371 \times 10^6 m$.*

   Solution: $P(r) = \frac{2\pi}{3} \rho_0^2 \mathcal{G} \left(R^2 - r^2 \right)$

The gravity field of a spherically symmetric density distribution is identical to the field of an equivalent point-mass. (see problem 16 for the spatial case of a uniform density distribution). This can be formulated as follows,

$$g(r) = \frac{Gm(r)}{r^2}, \qquad (17.34)$$

with

$$m(r) = \int_{V(r)} \rho dV = \int_0^r \rho(r') 4\pi r'^2 dr' \qquad (17.35)$$

Here $m(r)$ is the mass inside a sphere of radius r and $g(r)$ is the corresponding magnitude of the gravity acceleration. For the corresponding gravity potential this implies, with $\int_r^\infty \frac{dU}{dr'} dr' = U(\infty) - U(r) = -U(r)$,

$$U(r) = -\int_r^\infty \frac{dU}{dr'} dr' = \int_r^\infty g_r(r') dr' = \int_r^\infty -g(r') dr' = -\int_r^\infty \frac{Gm(r')}{r'^2} dr' \qquad (17.36)$$

where the radial vector component g_r has been expressed in the vector length g as $g_r = \mathbf{g} \cdot \mathbf{e}_r = -g$.

To derive (17.34), the potential field at the radial coordinate r can be split in contributions originating from an internal- and external density distribution $U(r) = U_i(r) + U_e(r)$. With corresponding pairs, $U_i \leftrightarrow \rho_i$, and $U_e \leftrightarrow \rho_e$, where $\rho_e(r') = 0$, $r' \leq r$, and $\rho_e(r') = \rho(r')$, $r' > r$. This follows from the linearity of the governing Poisson equation.

The field generated by the internal mass distribution is obtained by integrating the corresponding Poisson equation in spherical coordinates,

$$\frac{1}{r'^2}\frac{d}{dr'}r'^2\frac{dU_i}{dr'} = 4\pi G\rho_i \tag{17.37}$$

$$\int_0^r \frac{d}{dr'}\left(r'^2\frac{dU_i}{dr'}\right)dr' = \int_0^r 4\pi G\rho_i r'^2 dr' \tag{17.38}$$

The left hand side becomes simply $r^2\frac{dU_i}{dr}$ so the radial component of the gravity acceleration becomes

$$g_r(r) = -\frac{dU_i}{dr} = -\frac{1}{r^2}\int_0^r 4\pi G\rho_i r'^2 dr' \tag{17.39}$$

Furthermore the acceleration field g_e from the external mass distribution ρ_e for internal evaluation points $r' < r$ is zero. The corresponding gravity potential U_e is uniform, which follows from the relevant Poisson equation, in spherical coordinates for a spherically symmetric mass distribution,

$$\frac{1}{r'^2}\frac{d}{dr'}r'^2\frac{dU_e}{dr'} = 4\pi G\rho_e = 0 \;\rightarrow\; r'^2\frac{dU_e}{dr} = A \;\rightarrow\; g_e(r') = -\frac{dU_e}{dr'} = -\frac{A}{r'^2} \tag{17.40}$$

A non-singular field requires $A = 0$, $g_e(r') = 0$, $r' \leq 0$ and,

$$\frac{dU_e}{dr'} = 0 \;\rightarrow\; U_e(r') = B,\; r' \leq r \tag{17.41}$$

Looking back at the PREM model, it is a radial model $\rho(r)$ so can compute g_r at the surface of the Earth. The code is available in /images/prem/ and uses a somewhat naive numerical quadrature of Eq. (17.39):
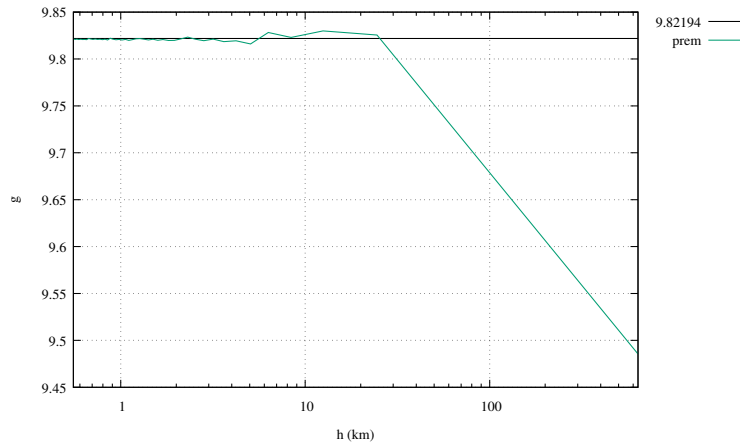
Radial gravity component measured at the surface of the Earth as a function of the cell size used for the integration.

For very high resolutions we do recover the analytical value as computed in Section 26.1.

problem: 17. *Verify that 17.36, applied to the special case of a homogeneous sphere of density ρ_0, lead to the same expression for the internal and external potential and acceleration field as given in problem 16.*

For a two-parameter spherically symmetric planet model consisting of a uniform core and mantle with radius R_c and R_m and contrasting densities ρ_c and ρ_m, the gravity field can also be determined by solving the Poisson equation for the particular density distribution and determination of the integration constants from the boundary conditions. However in this case the formula (17.34) are more convenient to obtain expressions for the gravity field.

---

problem: 18. *Derive expressions for the gravity acceleration and internal pressure distribution for the two-parameter model*

$$\rho(r) = \begin{cases} \rho_c, & r < R_c \\ \rho_m, & R_c < r \le R \\ \rho_e = 0, & r > R \end{cases} \quad , \quad g(r) = \begin{cases} g_c, & r < R_c \\ g_m, & R_c < r \le R \\ g_e, & r > R \end{cases} \quad , \quad P(r) = \begin{cases} P_c, & r < R_c \\ P_m, & r \ge R_c \end{cases}$$

(17.42)

using (17.34) and (17.9). See also (17.48).

Answer:

$$g_c(r) = \frac{4\pi}{3}G\rho_c r, \quad g_m(r) = \frac{G}{r^2}\left\{\frac{4\pi}{3}\rho_m\left(r^3 - R_c^3\right) + M_c\right\}, \quad g_e(r) = \frac{G}{r^2}\left(M_m + M_c\right) \qquad (17.43)$$

$$M_c = \frac{4\pi}{3}R_c^3\rho_c \ , \quad M_m = \frac{4\pi}{3}\rho_m\left(R^3 - R_c^3\right) \qquad (17.44)$$

$$P_c(r) = P_m(R_c) + \frac{2\pi}{3}G\rho_c^2\left(R_c^2 - r^2\right) \qquad (17.45)$$

$$P_m(r) = \frac{2\pi}{3}G\rho_m^2\left\{R_m^2 - r^2 + 2\left(\frac{\rho_c}{\rho_m} - 1\right)R_c^3\left(\frac{1}{r} - \frac{1}{R_m}\right)\right\} \qquad (17.46)$$

17.5.2 The pressure effect on density

In the previous sections we considered the gravity field of a given mass distribution. For self-gravitating planets of sufficient size **the local density depends on the pressure, through selfcompression** i.e. the compression of the material caused by the planets own gravity field. As we have seen in previous sections the lithostatic pressure depends on the gravity field and the density distribution. It follows that the determination of the density, gravity and pressure are coupled problems that must be solved simultaneously and can not be solved separately. Here we will consider the solution of such coupled problems.

From observations of the average density of surface rocks of some $2.7 \cdot 10^3$ kg/m^3 and the known mean density of the Earth $5.5 \cdot 10^3$ kg/m^3, it follows that the surface density is less than half the mean Earth value. **The difference between both density values suggests a density increase in the interior which could be related either to different composition at depth**, for example corresponding to a dense metallic core, **and/or the effect of selfcompression in an otherwise homogeneous planet**. Solid state phase transitions of mantle material due to increasing pressure can also explain part of the high mean density value.

From the nineteenth century on, models of the internal density distribution of the earth have been investigated. These models have in common that the radial density distribution is parameterized in a simple way with a small number of parameters, typically two, which are then adjusted to the known data such as the surface density and the Earth's total mass or moment of inertia.

In the following the relation between density, gravity and pressure in a self-gravitating planet will be investigated in a more self consistent way.

For a spherically symmetric density distribution the corresponding magnitude of the gravity accel-

eration vector is given by (17.39),

$$g(r) = |\boldsymbol{g}(\boldsymbol{r})| = |g_r(r)| = \frac{4\pi G}{r^2} \int_0^r \rho(r')r'^2 dr' = \frac{Gm(r)}{r^2} \tag{17.47}$$

where $m(r)$ is the mass of a sphere of radius r and $\rho(r)$ is the corresponding radial density profile.

> **problem: 19.** *Use (17.47) to show that it is not possible to derive a unique radial mass distribution of a spherically symmetric planet from the observed surface value of the gravity field alone. This can be verified by showing that multiple density profiles exist that produce the same surface gravity. To illustrate this sketch a schematic internal radial profile of the gravity acceleration in a comparison of two spherically symmetric planets of identical mass M and radius R. The first one is a homogeneous planet with density ρ_0 and the second one is a differentiated planet with a uniform high density core $\rho_c = \rho_0 + \delta\rho$ and less dense mantle $\rho_m = \rho_0 - \delta\rho$. Verify that these assumptions correspond to this special case with volume fraction of the core $\phi_c = 1/2$.*

From the above results the lithostatic pressure distribution can be obtained by substitution for the gravity acceleration and integrating the pressure gradient $dP/dr = -\rho g$. Assuming a zero pressure value at the surface this results in,

$$P(r) = \int_r^R \rho(r')g(r')dr' = 4\pi G \int_r^R \rho(r') \left\{ \frac{1}{r'^2} \int_0^{r'} \rho(r'')r''^2 dr'' \right\} dr' \tag{17.48}$$

The pressure in the Earth's interior reaches values over 350 GPa as shown in Fig. **??**. For such high pressure values the effect of self-compression on the density is significant. In the following this effect is further explored.

The bulkmodulus An isotropic linear elastic solid can be described by two independent elasticity parameters, for instance the Lamé parameters λ and μ [8]. The bulkmodulus can be expressed in the Lamé parameters as, $K = \lambda + \frac{2}{3}\mu$. The bulkmodulus K and the shearmodulus μ are the most commonly used parameters to specify the elastic parameters of Earth materials.

The bulk modulus K of a substance measures the substance's resistance to uniform compression. It is defined as the ratio of the infinitesimal pressure increase to the resulting relative decrease of the volume. Its SI unit is the Pascal, and its dimensional form is $M^1 L^{-1} T^{-2}$.
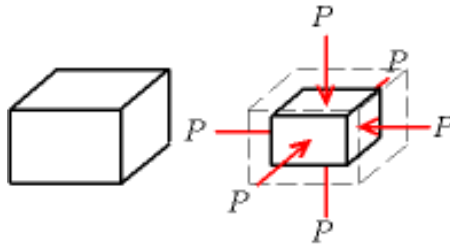
The incompressibility K, or bulkmodulus , is defined as,

$$\frac{1}{K} = \frac{1}{\rho}\frac{d\rho}{dP} \tag{17.49}$$

By substitution of $dP = -\rho g dr$ in (17.49) we derive a differential equation for the density profile of a compressible planet model,

$$\frac{1}{K} = \frac{-1}{\rho^2 g}\frac{d\rho}{dr} \Rightarrow \frac{d\rho}{dr} = -\frac{\rho^2 g}{K} \tag{17.50}$$

---

[8]https://en.wikipedia.org/wiki/Elastic_modulus

Parameterization of the bulkmodulus The radial density distribution for a selfcompressing planet can be obtained from (17.50) once the bulkmodulus K is known. We will first consider simple cases where K is either a uniform constant or it is parameterized in terms of the density.

---

problem: 20. *Assume both K and g in (17.50) to be uniform in the mantle and derive the following density profile,*

$$\rho(z) = \frac{\rho_0}{1 - \frac{\rho_0 g z}{K}} \tag{17.51}$$

where $z = R - r$ is the depth coordinate and $\rho_0 = \rho(0)$ is the surface density value.

- *Compute the depth z_1 where the expression (17.51) becomes singular, i.e. $\rho \to \infty$, suggesting infinite compression of the material. To do this assume Earth(mantle)-like values of the incompressibility, $K = 400\text{GPa}$ (see Fig.??) and the surface density $\rho_0 = 3 \cdot 10^3 \text{ kg/m}^3$.*

- *Now consider a simplified model of a large rocky exoplanet of Earth-like composition with $M = 8M_\oplus$ and $R = 1.5R_\oplus$. Assume uniform gravity (adapted for the given M, R) and uniform incompressibility K. Do you now find the singular depth z_1 within the depth range of the planet? Comment on the assumption of a uniform gravity field in view of the models presented in section 17.5.1.*

---

problem: 21. *The result of problem 20 gives the density depth distribution for the model with constant properties. The resulting expression (17.51) also contains the uniform gravity acceleration. A more fundamental relation between density and pressure, not including gravity, can be derived for this model with constant material property K as an equation of state (EOS) for the density.*

Derive from the definition of the bulkmodulus (17.49) the following logarithmic EOS for the density in terms of the static pressure,

$$P = \ln\left(\left(\frac{\rho}{\rho_0}\right)^K\right) \tag{17.52}$$

Show that the above EOS (17.52) can be inverted to obtain an explicit expression for density as a function of pressure.

The singular behavior in the density model of problem 20 is a result of the assumed uniform g and K in (17.50). While g is reasonably constant with depth in the mantle, as illustrated in Fig. ??, K is not. The incompressibility increases with increasing depth/pressure and as a result the compression remains finite for earth-like conditions. The incompressibility can be expressed in the density and the seismic wave velocities, $v_p = \sqrt{(\lambda + 2\mu)/\rho}$, $v_s = \sqrt{\mu/\rho}$. With $K = \lambda + \frac{2}{3}\mu$ this becomes $K = \rho(v_p^2 - 4/3v_s^2)$. A radial profile $K(P(r))$ can therefore be derived, from the seismic velocities determined from inversion of traveltime tables of longitudinal and shearwave seismic arrivals.

The $K(P(r))$ profile derived from the PREM model of Dziewonski and Anderson (1981) appears to be roughly linear as shown in the following figure:
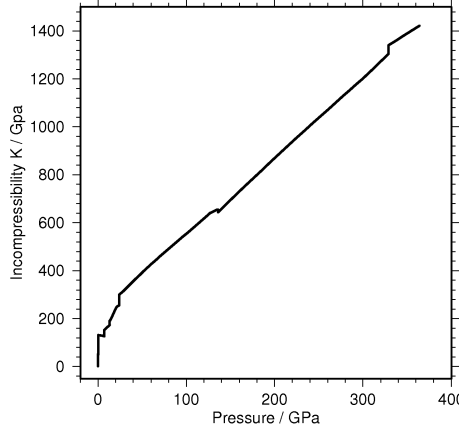
Incompressibility profile derived from the PREM model.

A linear relation between bulkmodulus and pressure as suggested by this figure is also obtained using the following power law parameterization for the bulkmodulus in terms of the density $K(\rho)$.

$$K = C\rho^n \quad \Rightarrow \ln(K) = \ln(C) + n\ln(\rho) \quad \Rightarrow n = \frac{d\ln(K)}{d\ln(\rho)} = \frac{dK}{dP} = K_0' \tag{17.53}$$

where C is a constant. The constant pressure derivative in this model implies a linear pressure relation $K(P) = K_0 + K_0'P$. This appears to approximate the distribution of K in particular in the lower mantle as determined from seismological data in the PREM model. $K_0' \approx 4$ for the magnesium-iron sillicates $(Mg, Fe)SiO_3$ (perovskite) and dense oxides $(Mg, Fe)O$ (wüstite), representative for the earth's deep mantle.

The Murnaghan e.o.s. An equation of state directly relating the density or specific volume, $V = 1/\rho$, to pressure can be derived from such an 'ansatz' of a linear pressure dependence $K = K_0 + K_0'P$ as shown in the following,

$$\frac{1}{\rho}\frac{d\rho}{dP} = \frac{1}{K} \quad \rightarrow \quad \frac{1}{V}\frac{dV}{dP} = -\frac{1}{K} \quad \rightarrow \quad dP = -(K_0 + K_0'P)\frac{1}{V}dV \tag{17.54}$$

$$\int_0^P \frac{dP'}{K_0 + K_0'P'} = -\int_{V_0}^V \frac{1}{V'}dV' = \int_V^{V_0} \frac{1}{V'}dV' = \ln\left(\frac{V_0}{V}\right) \tag{17.55}$$

Substitution in the integral over pressure of $K_0 + K_0'P' = x$, $dx = K_0'dP'$ gives,

$$\int_{x_0=K_0}^{x_P=K_0+K_0'P} \frac{1}{K_0'}\frac{dx}{x} = \frac{1}{K_0'}\ln\left(\frac{K_0+K_0'P}{K_0}\right) = \ln\left(\frac{V_0}{V}\right) \tag{17.56}$$

$$1 + \frac{K_0'P}{K_0} = \left(\frac{V_0}{V}\right)^{K_0'} \quad \rightarrow \quad P = \frac{K_0}{K_0'}\left(\left(\frac{V_0}{V}\right)^{K_0'} - 1\right) \tag{17.57}$$

This relation is known as the Murnaghan equation of state (EOS).

The Murnaghan equation of state is a relationship between the volume of a body and the pressure to which it is subjected. This is one of many state equations that have been used in earth sciences and shock physics to model the behavior of matter under conditions of high pressure. It owes its name to Francis D. Murnaghan who proposed it in 1944 to reflect material behavior under a pressure range as wide as possible to reflect an experimentally established fact: **the more a solid is compressed, the more difficult it is to compress further**.

The Murnaghan equation is derived, under certain assumptions, from the equations of continuum mechanics. It involves two adjustable parameters: the modulus of incompressibility K_0 and its

846

first derivative with respect to the pressure, K_0', both measured at ambient pressure. In general, these coefficients are determined by a regression on experimentally obtained values of volume V as a function of the pressure P. These experimental data can be obtained by X-ray diffraction or by shock tests. Regression can also be performed on the values of the energy as a function of the volume obtained from ab-initio and molecular dynamics calculations.

---

problem: 22. *Derive an explicit expression for the pressure dependent density from the Murnaghan equation of state* (17.57).
Answer:

$$\rho(P) = \rho_0 \left(\frac{K_0' P}{K_0} + 1 \right)^{1/K_0'} \tag{17.58}$$

---

problem: 23. *In problem 20 we have seen that a simple model with uniform incompressibility and gravity $K = K_0$ and $g = g_0$ leads to physically impossible solutions. In a refined version of this model, applied to the Earth's mantle, $g = g_0$ is maintained (compare Fig.??), and K is parameterized using the powerlaw relation* (17.53).
Derive the following density profile for the model corresponding to (17.53).

$$\rho(r) = \rho_0 \left(1 + (n-1) \frac{\rho_0 g_0 z}{K_0} \right)^{\frac{1}{n-1}} \tag{17.59}$$

where $z = R - r$ is the depth coordinate and the 0 subscript refers to zero pressure conditions. Note that the singularity for $\rho_0 g_0 z / K_0 = 1$ in problem 20 is absent in this model.

A more widely used and more accurate EOS for a higher pressure range is the equation derived by Birch (1952) from a consideration of elastic strain energy, known as the Birch-Murnaghan EOS (Poirier, 2000).

In other cases than the special simplified cases discussed above, in particular in problems 20 and 23, the gravity acceleration varies also with depth. Also more accurate equations of state may be necessary for very high pressure, encountered in the deep interior of large (exo)planets, that result in large compression. Such models can be formulated in a more general way by the following coupled set of equations for pressure, gravity and density.

$$\frac{dP}{dr} = -\rho g \tag{17.60}$$

$$g(r) = \frac{Gm(r)}{r^2} \tag{17.61}$$

$$F(\rho, P, T) = 0 \tag{17.62}$$

where the radial mass distribution $m(r)$ is defined as in (??). A model based on (17.60), (17.61), and (17.62) can be constructed for the internal structure (density, gravity, pressure) of a planet of given mass M and composition, i.e. with given parameters of the EOS (17.62) such as ρ_0, K_0, K_0' in the Murnaghan EOS (17.57). Consider the application of such a model to a planet for which only the planet mass M is known. [9] Assume a homogeneous terrestrial (rocky) planet without a distinct metallic core. Assuming an earth-mantle like composition, representative values of the EOS parameters can be used, to solve the coupled model equations in the following iterative scheme.

1. Define a grid along the radial coordinate $r_i, i = 1, \ldots, N$, $r_1 = 0$. This grid defines a subdivision of the interior in $N - 1$ concentric layers and must be chosen large enough, i.e. $r_N > R$.

---

[9]Such models can be applied to exoplanets that are recently being discovered https://en.wikipedia.org/wiki/Methods_of_detecting_exoplanets . For some of these planets, detected from radial velocity variations of the star, only the planet mass M is known.

2. Choose an initial estimate of the central pressure $P^{(1)}(0)$.

3. In a loop over the internal layers, starting upward from the centre, first compute the pressure decrement over the layer from (17.60). This is then used to obtain the pressure at the next grid point and corresponding density from the EOS (17.62). From the computed density the corresponding mass distribution $m(r_i)$ and gravity $g(r_i)$ (17.61) follow.

4. The layer iteration in the previous item is stopped when a zero pressure value has been reached. The radial level reached this way now defines the next approximation of the planetary radius $R^{(j)}$ and $M^{(j)} = m(R^{(j)})$ is a new approximation of the planet mass M.

5. From the total mass defect $\Delta M^{(j)} = M^{(j)} - M$ a correction to the central pressure is computed as $\Delta P^{(j)}$, (problem 24). In the next iteration the radial integration is repeated from item 3 with an updated central pressure $P^{(j+1)}(0) = P^{(j)}(0) + \Delta P^{(j)}$ and this iterative procedure is repeated until convergence is reached, i.e. until $|\Delta M^{(j)}|/M$ drops below a specified tolerance value.

---

problem: 24. *A correction for the central pressure in item 4 can be estimated by distributing the mass defect $\Delta M^{(j)}$ over a spherical shell of thickness $\Delta R^{(j)}$, positioned at the surface, and computing an approximate pressure $\Delta P^{(j)}$ at the bottom of this shell.*
Derive the following expression for the thickness of this spherical shell,

$$\frac{\Delta R^{(j)}}{R^{(j)}} = \left(\frac{\Delta M^{(j)}}{M^{*(j)}} + 1 \right)^{1/3} - 1 \tag{17.63}$$

Where $M^{*(j)} = \frac{4\pi}{3}\rho(R^{(j)})R^{(j)3}$.
The correction for the central pressure is then defined as, $\Delta P^{(j)} = \rho(R^{(j)})g(R^{(j)})\Delta R^{(j)}$.

17.5.3 Adiabatic density distribution

In the previous section density models were based on assumptions about the parameterization of the bulkmodulus K. The density model of Williamson and Adams (1923), (Hemley, 2006) does not depend on a parameterized K. Instead it is defined in terms of the seismic wave velocities v_p and v_s that can be determined from inversion of seismological traveltime data as $K/\rho = v_p^2 - 4/3v_s^2$.

The W-A model can be derived from thermodynamic principles for a homogeneous self-compressing layer which is in an adiabatic state. The bulkmodulus applied in this model is expressed in the seismic wave velocities which in turn depend on the elasticity parameters and the density. The elastic deformation proces in seismic wave propagation occurs on a relatively short time scale (seconds-minutes) compared to the characteristic time scale of conductive heat transport in solids (see **??**). Therefore (diffusive) heat exchange can be neglected and adiabatic conditions apply in seismic wave propagation. This implies that the elasticity parameters determined from seismic data, including the bulkmodulus K pertain to adiabatic conditions (see also Appendix **??**).

Other processes such as convective mantle flow that occur on a much longer time scale may take place under more general (non-adiabatic) conditions.

In section **??** on the thermal state of the Earth it is shown that adiabatic conditions hold for the interior of a fluid layer when heat transport is dominated by advection and heat diffusion by conduction/radiation plays a minor role. Assuming the Earth's mantle to be in a state of vigorous thermal convection it also follows that the average temperature profile, the geotherm, corresponds to an adiabatic distribution.

In general the density differential can be written as,

$$d\rho = \left(\frac{\partial \rho}{\partial P}\right)_S dP + \left(\frac{\partial \rho}{\partial S}\right)_P dS \tag{17.64}$$

where the differential of the entropy S is dropped in case of adiabatic conditions and the pressure derivative is written in terms of the adiabatic bulkmodulus K_S defined in (17.49), $1/K_s = (\partial \rho/\partial P)_S/\rho$.

---

problem: 25. *Derive the Williamson-Adams equation for a homogeneous adiabatic layer from the density differential (17.64) and assumption of isentropic (adiabatic) conditions with $dS \equiv 0$,*

$$\frac{d\rho}{dr} = -\frac{\rho^2 g}{K_S} \tag{17.65}$$

The density solution of the W-A equation can be expressed in terms of the seismic parameter $\Phi = K_S/\rho$ which in turn can be obtained from seismic velocity models: $\Phi = v_p^2 - \frac{4}{3}v_s^2$ for P and S waves. $\sqrt{\Phi} = \sqrt{K_S/\rho}$ is known as the bulkvelocity. For a given bulkvelocity profile, obtained from seismic observations, the W-A density profile is derived from (17.65) as,

$$\ln\left(\frac{\rho(r)}{\rho(R)}\right) = \int_r^R \Phi^{-1}(r')g(r')dr' \tag{17.66}$$

---

problem: 26. *Derive (17.66) by integration of the W-A equation (17.65).*

In (17.66) the gravity acceleration g depends on the density distribution $\rho(r)$ in the lefthand side. Therefore the density profile can not be simply obtained from a seismologically determined $\Phi(r)$ profile and a single evaluation of the integral in (17.66). The expression represents an integral equation that can be solved iteratively as specified in problem 27.

---

problem: 27. *Assume that a seismic parameter profile for the mantle $\Phi(r)$, obtained from seismic travel times, is available. Investigate how (17.66) can be used to compute a sequence of mantle density profiles $\rho^{(j)}(r), j = 1, 2, \ldots$ in an iterative procedure, by succesive substitution. How would you define a starting profile $\rho^{(1)}(r)$ for this iterative procedure?*
Hint: Substitute the density profile for iteration number j in the gravity acceleration in the righthand side of (17.66) for the computation of an updated profile $j + 1$. This is an example of a general solution strategy for non-linear problems known as 'succesive substitution' or Picard iteration.

Williamson and Adams (1923) [1463] used the iterative scheme in problem 27 to test the hypothesis that the mass concentration towards the Earth's centre is completely explained by compression of a homogeneous self-gravitating sphere. They showed that integrating (17.66) from a surface value of $3.3 \cdot 10^3$ kg/m^3 results in unrealistically high density values for depths greater than the core-mantle boundary. This way they concluded that an inhomogeneous earth with a dense, compositionally distinct core, probably iron-nickle, was required by the observations. The necessary multiple integrals in the evaluation of (17.66) had to be computed by means of graphical approximation methods in 1923, several decades before the advent of electronic computers.

In a later analysis Bullen (1936) showed that the assumption of a homogeneous selfcompressing mantle described by the W-A equation, and a chemically distinct dense core, leads to unrealistically high values of the moment of inertia for the core $I_c = fM_cR_c^2$, with a prefactor value $f \sim 0.57$ greater than the value of a core with uniform density, 0.4. Since this would imply a density decrease towards the centre Bullen concluded that the applicability of the W-A model for the whole mantle

can not be maintained and that instead a distinct mantle transition layer, labeled C-layer, must be included between the upper and lower mantle proper, related to transitions in mineral phase and/or composition (Bullen, 1975).

---

problem: 28.

1. *Derive the following equation for the temperature distribution of a W-A layer (see Appendix ??),*

$$\frac{dT}{dr} = -\frac{\alpha g}{c_P} T \tag{17.67}$$

   where α and c_P are the thermal expansion coefficient and the specific heat at constant pressure.
   Hint: Use the differential for the entropy,

$$dS = \left(\frac{\partial S}{\partial T}\right)_P dT + \left(\frac{\partial S}{\partial P}\right)_T dP \tag{17.68}$$

   and the thermodynamic relations: $(\partial S/\partial T)_P = c_P/T$ and $(\partial S/\partial P)_T = -\alpha/\rho$.

2. *Derive the expression for the temperature profile for an adiabatic layer, sometimes referred to as the 'adiabat', by solving equation (17.67),*

$$T(r) = T(R) \exp\left(\int_r^R \frac{\alpha g}{c_P} \, dr'\right) \tag{17.69}$$

   The temperature extrapolated to the surface, $T_P = T(R)$ is known as the potential temperature of the layer. The quantity $H_T = (\alpha g/c_P)^{-1}$ is known as the thermal scale height of the layer.

3. *Derive an expression from (17.69) for the special case with a constant value of the scale height parameter.*

The W-A equation for the density of an adiabatic layer can be generalized introducing the Bullen parameter η which is used as a measure of the departure of the actual density/temperature profile from an adiabat. This is done by writing,

$$\eta(r) = -\frac{\Phi}{\rho g} \frac{d\rho}{dr} \tag{17.70}$$

where $\eta(r)$ has been substituted for the constant value ($\equiv 1$) in the W-A equation.

Current density models

The concept of an adiabatic layer was essential when no independent determinations for the density distribution were available and the W-A equation was used to compute $\rho(r)$ for given values of the seismic parameter $\Phi(r)$ determined from seismological observations (Bullen, 1975).

During the 1970s a radial density distribution has been obtained for the Earth from inversion of seismological observations, incorporating spectral analysis of the Earth's eigenvibrations, under the constraints of the given values for M and I. This, together with seismic velocities determined from bodywave traveltimes and surfacewave dispersion, has resulted in the Preliminary Reference Earth Model (PREM), (Dziewonski and Anderson, 1981 [380]).

Since $\rho(r)$ can be determined from analysis of the earth's normal modes (radial eigenvibrations) the 'adiabaticity' of the mantle is no longer assumed.

The degree of 'adiabaticity' is used in numerical modelling experiments as a diagnostic for the dynamic state - where a high degree of adiabaticity indicates vigorous thermal convection and predominantly convective heat transport (van den Berg and Yuen, 1998, [1395] Matyska and Yuen, 2000, [892] Bunge *et al.* , 2001).

Usually the outcome of such experiments shows that the upper and lower mantle separately are approximately adiabatic - away from boundary layers were conductive transport dominates. In recent years models of the deep lower mantle have become popular were a compositionally distinct dense layer occupies the bottom 30% (roughly) of the lower mantle (Kellog *et al.* (1999) [733], Albarede and van der Hilst (2002) [4]).

17.5.4   Earth's chemical composition

For a complete description of the Earth's interior we need to know it's chemical composition, temperature and pressure. In section 17.4 the pressure is expressed in the density distribution and the related internal gravity field. Once the internal pressure distribution is known, sharp transitions or discontinuities in the material properties, like the seismic velocities v_p, v_s and the density in the PREM model, can be identified with mineral phase transitions and as such they can be related to the mineral (P, T) phase diagram of candidate mantle silicate materials in order to estimate the temperature in the Earth's interior. Such phase diagrams are determined from experimental (HPT) and theoretical work in mineral physics.

What do we know about Earth's bulk chemical composition? Candidate mantle materials have been defined based on cosmochemical and petrological considerations. Models of the chemical composition of the Earth are commonly based on the hypothesis that the planet was formed in a multi-stage accretion proces from material that condensated from the original solar nebula approximately 4.6 billion years ago at the time of formation of the solar system. The chemical composition of chondritic meteorites, in particular the carbonaceous chondrites (CI type) (McBride and Gilmour (2003) [898]) show a strong correlation with the composition of the outer layer of the sun (photosphere), determined from spectral analysis of the solar light, as illustrated in the following figure:
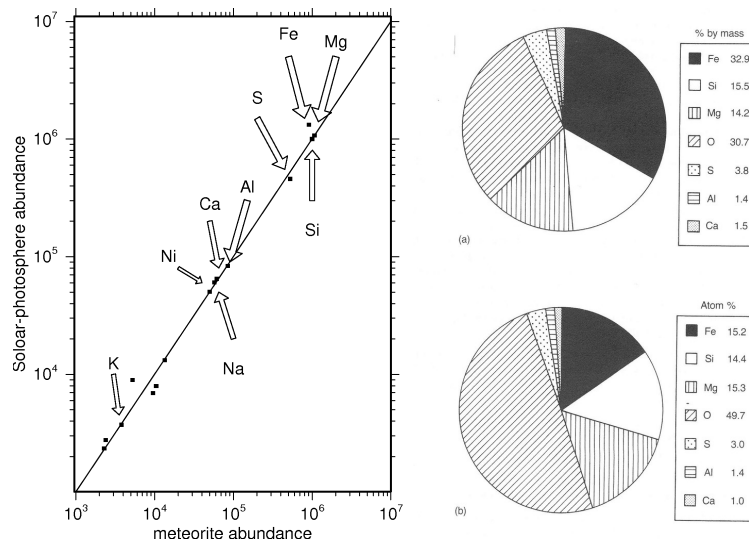
Left: Element abundance (normalized with $Si = 10^6$), of the solar shallow photosphere compared to chondritic meteorites (Anders & Grevesse (1989) [18]).

Right: amounts of Earth's major elements assuming a chondritic composition (Brown & Musset (1993) [168]).

The solar-chondritic data in the lefthand frame show that Mg, Fe and Si are by far the most abundant (non-volatile) elements. According to the chondritic Earth hypothesis a similar abundance can be expected for the bulk-earth. This is illustrated in the righthand pie diagrams. Note the large proportion of oxygen, bound in oxides. In most crust-mantle rocks S is less abundant than Al or Ca. This is usually explained by assuming that S is relatively volatile and also 'siderophile', meaning

that a significant fraction may have ended up in the iron-nickle core during an early core-mantle differentiation.

The chondritic meteorites are thought to be representive of the undifferentiated material condensated from the solar nebula.

Around 1960 a model chemical composition for the bulk of the Earth's mantle, coined pyrolite, was introduced by Ringwood (see (Ringwood, 1975) and original references therein). This is still used as a reference model. The pyrolitic composition is associated with the main upper mantle rock type peridotite that is brought to the Earth's surface in small fragments included in volcanic rocks (xenoliths) and also in larger, kilometer sized, fragments in so called peridotite bodies (Spengler *et al.*, 2006). The pyrolitic composition of the upper mantle rocks is also strongly correlated with the composition of chondritic meteorites, in agreement with the hypothesis of a chondritic origin of the Earth.

Mantle peridotites are found with different degrees of depletion (mass fraction lost) by partial melting. More depleted material is denoted as harzburgite and the relatively undepleted peridotite is known as lherzolite. During progressive partial melting the mineral composition of the residual rock material, a mineral assemblage consisting of olivine, pyroxene and garnet, shifts towards the olivine composition. The olivine enriched harzburgitic residue appears to be the chemical complement of the basaltic melt product, with respect to the original lherzolitic mantle source rock. This depletion relation, between oceanic and continental crust on the one hand and peridotitic mantle rock on the other, is reflected in the element abundance of crust and mantle rocks, illustrated in the following figure:
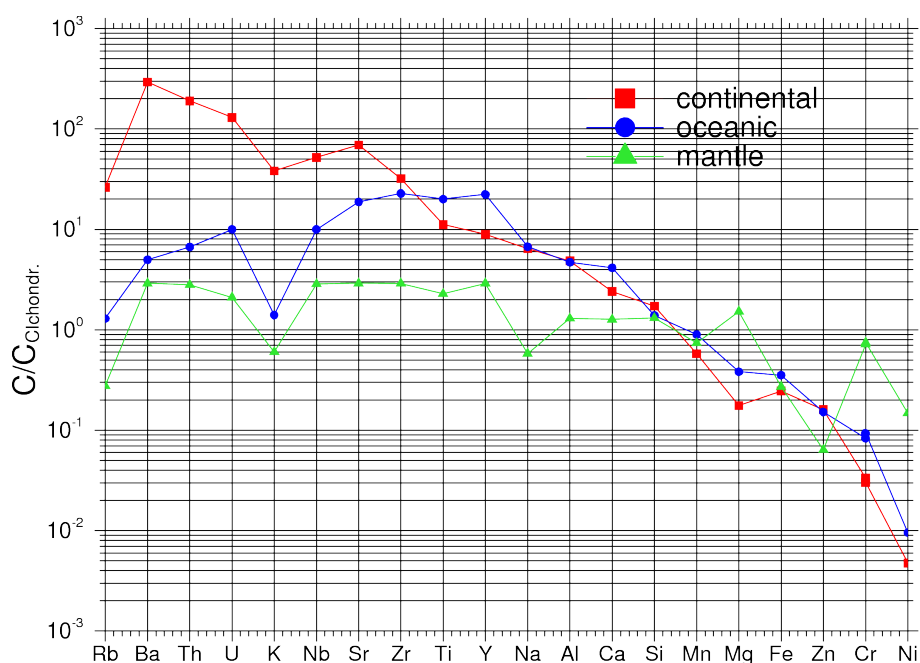
Chemical abundance of crustal and mantle rocks, normalized with respect to CI chondritic values. Data from (McBride and Gilmour, 2003 [898]).

This figure shows abundance ratio's relative to the CI-chondritic composition. The curve for mantle rock appears to be relatively close to the chondritic composition, whereas the crustal material is enriched with respect to the mantle in most elements shown.

A notable exception to this crustal enrichment is found for magnesium which appears to be enriched in average mantle peridotite. This is in agreement with the previous observation that the olivine/pyroxene content ratio of the residual increases with the degree of partial melting. Magnesium content increases with the olivine (forsterite Mg_2SiO_4)/pyroxene $MgSiO_3$ ratio.

An other observation that can be made from the figure above is the apparent depletion of the siderophile elements Fe and Ni, both in crust and mantle material, with respect to the chondritic composition. This is usually explained by the formation of a liquid Fe, Ni rich metal core of the

Earth during the first few million years of the accretion proces, in the early solar system. During this event the molten liquid metal would have differentiated from the silicate mantle, leaving the mantle depleted in siderophile (iron loving) elements.

Core formation is also sometimes used as an explanation of the apparent K (potasium) depletion of both mantle and crust with respect to chondrites. In this explanation K is disolved in liquid iron in significant quantity at high pressure and temperature (Rama Murthy *et al.*, 2003). An alternative explanation for the Earth's K depletion is an escape of K due to significant volatilization during the planetary accretion proces.

> **problem: 29.** *From Figure* **??** *it can be concluded that the Earth's mantle and crust lost roughly 2/3 of its original iron content corresponding to a chondritic composition. Verify how this iron-depletion of crust and mantle could be explained by differentiation of the Earth's mostly-iron core. Use the following data in your argument: a) The mass fraction of the core $X_c = M_c/M_\oplus = 0.315$. b) The Fe mass fraction $X_{mFe} \sim 10\%$ of the pyrolitic mantle, c) The mass fraction of lighter elements in the core - (S, Si, O) amounts to about 20%. d) The Fe mass fraction of the bulk Earth $X_{\oplus Fe} \sim 33\%$ (Fig.* **??***)*

17.5.5 Phase transitions as anchor points of the geotherm

Major phase boundaries in the Earth's mantle and core have been identified with sharp transitions in the seismic wave velocities and the density distribution of the PREM model.

The depth distribution of the mineral composition for a pyrolitic mantle model is shown in the following figure:
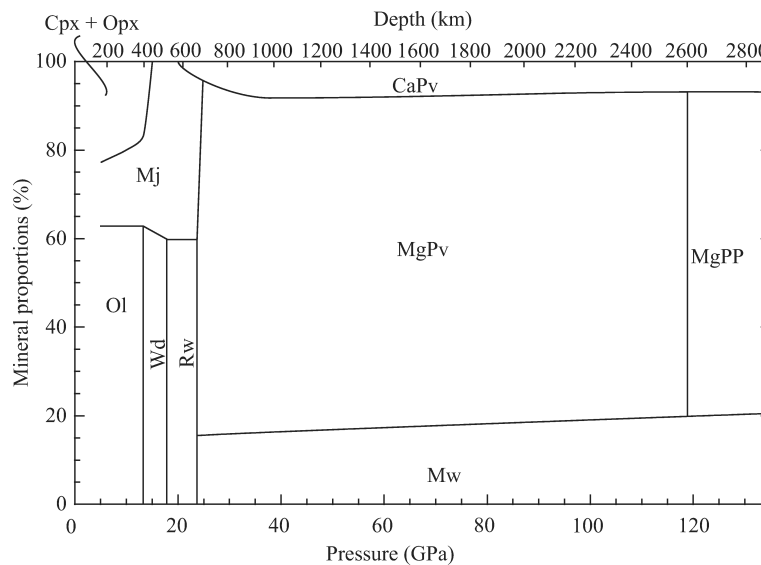
Pressure/depth distribution of mineral assemblage for a pyrolitic mantle model. Cpx: clinopyroxene, Opx: orthopyroxene, Mj: majorite garnet, Ol: olivine, Wd: wadsleyite, Rw: ringwoodite, CaPv: $CaSiO_3$ perovskite MgPv: $MgSiO_3$-rich perovskite, MgPP: $MgSiO_3$-rich post-perovskite, Mw: magnesiowüstite.

(From: (Hirose, 2007))

This figure clearly illustrates the different mineral composition of the upper and lower mantle regions separated by the major phase boundary near 660 km depth ($\sim 24\,\mathrm{GPa}$), where the ringwoodite polymorph of olivine, $(\mathrm{Mg}, \mathrm{Fe})_2\mathrm{SiO}_4$, transforms (dissociates) into a mineral assemblage of perovskite, $(\mathrm{Mg}, \mathrm{Fe})\mathrm{SiO}_3$ and magnesiowüstite, $(\mathrm{Mg}, \mathrm{Fe})\mathrm{O}$.

For a given mantle composition, for instance for a pyrolitic mantle, the pressure-temperature mineral phase diagram can be determined for the relevant P, T range of the Earth's mantle by HPT experiments and mineral physics theory. A sharp transition at a pressure P_t in the PREM model can then be located at the corresponding pressure in the phase diagram by the intersection of the P_t isobar with the diagram phase boundaries. The (possibly multiple) intersection points

define the corresponding transition temperature T_t. The pressure-temperature point located in the phase diagram defines an 'anchor point' that constrains the geotherm. In this procedure the phase transition is used as a mantle/core thermometer.

This way several (P, T) 'anchor points' of the geotherm have been determined, related to the solid state phase transition near 660 km depth and the solid/liquid inner/outer core boundary at 1220 km from the Earth's centre.

The following figure from Boehler (1996) [113]) illustrates the determination of anchor points of the geotherm at the phase boundary near 660 km depth ($P_{660} = 24$GPa, $T_{660} = 1900 \pm 100$ K) and at the boundary between the outer and inner core at 5150 km depth, ($P_{ICB} = 330$GPa, $T_{ICB} = 4850 \pm 200$ K).
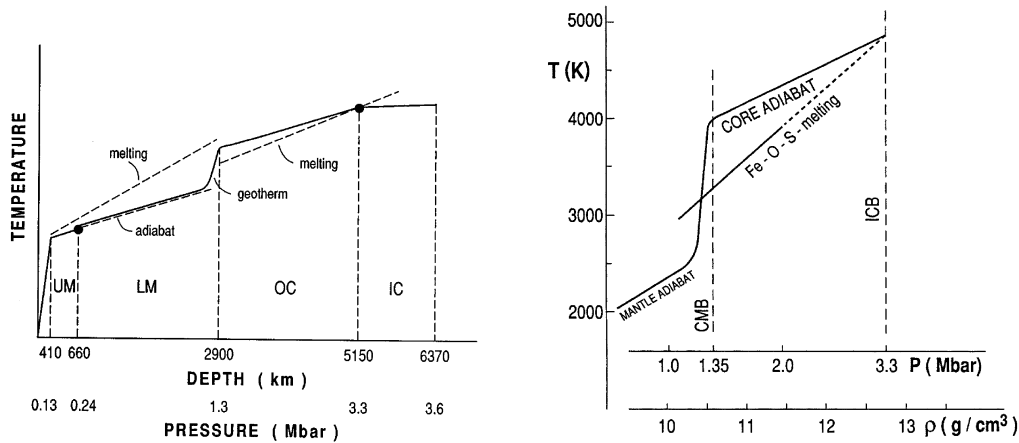
Schematic radial temperature distribution in the mantle and core, constrained by major phase transitions (Boehler, 1996), (UM-upper mantle, LM lower mantle, OC outer core, IC inner core). The temperature of the upper/lower mantle boundary is constrained by the γ-spinel to postspinel phase transition at 660 km depth. The temperature at the inner/outer core boundary at 5150 km depth (radius 1220 km) is constrained by the melting temperature of the hypothetical core 'Fe-O-S' alloy. The right hand frame shows a schematic core temperature distribution (geotherm) labeled 'CORE ADIABAT' in the liquid outer core versus pressure and the melting curve (liquidus) of the core 'Fe-O-S' alloy. (CMB core-mantle boundary, ICB inner core boundary). The ICB is determined by the intersection of the liquidus and the geotherm. During core cooling the ICB moves outward as the inner core grows by crystallisation.

Starting from these anchor points the temperature is then extrapolated from both sides to the core mantle boundary at 2900 km depth. For this temperature extrapolation assumptions have to be made about the dominant heat transport mechanism and in this case it is assumed that heat transport operates mainly through thermal convection. This will be further investigated in later sections dealing with heat transport in the Earth's mantle.

> **problem: 30.** *Estimate the temperature near the bottom of the mantle by adiabatic extrapolation of the temperature $T_{660} \sim 1900$K of the phase transition near 660 km depth, to the depth of the core mantle boundary, using the general expression for the adiabat in a homogeneous layer.*
> *Hints: apply the result of problem 28 and assume uniform values of the 'scale height parameter' $H_T = (\alpha g / c_P)^{-1}$, with $\alpha = 2 \cdot 10^{-5}$K^{-1}, $g = 10$ms^{-2}, $c_P = 1250$Jkg^{-1}K^{-1}. Further: approximate the adiabat by a linear depth function, in agreement with the schematic diagram of Boehler (1996) - see figure above -to obtain a uniform adiabatic temperature gradient.*

The 'head' of the extrapolated outer core adiabat is at a temperature of approximately 4000 K and the 'foot' of the lower mantle adiabat at approximately 2700 K. This result indicates a large temperature contrast of about 1300 K across the CMB.

How can such a large contrast be explained physically? As we will see later, this can be explained by interpreting the CMB as a boundary between two separately convecting fluid layers, each with a thermal boundary layer where the main heat transport mechanism shifts from convection in the interior of the fluid layers, to conduction near the boundary interface, where vertical convective transport vanishes with the flow velocity component normal to the boundary. Separately convecting

layers are in agreement with the large density contrast across the CMB where the density almost doubles, as illustrated in the PREM profile. The resulting strong temperature contrast across the CMB is consistent with a lower mantle in a state of vigorous thermal convection.

> **problem: 31.** *Explain why we can not turn this argument around and conclude from these indications for a strong temperature contrast at CMB that the mantle convects vigorously.*
> *Hint: Check Appendix ?? for the assumptions made for an adiabatic geotherm in the lower mantle.*

More recent developments, providing independent information, shed new light on the temperature distribution in the bottom layer of the lower mantle. A previously unknown mantle phase transition has been identified, in the main constituent magnesium-perovskite, to a ($\sim 1.5\%$) denser phase (post-perovskite) both in experimental HPT and theoretical (mineral physics) work at temperatures and pressure conditions corresponding to a region in the lowermost mantle close to the core-mantle boundary. This is illustrated in the figure hereafter showing experimental data points delineating the phase boundary.
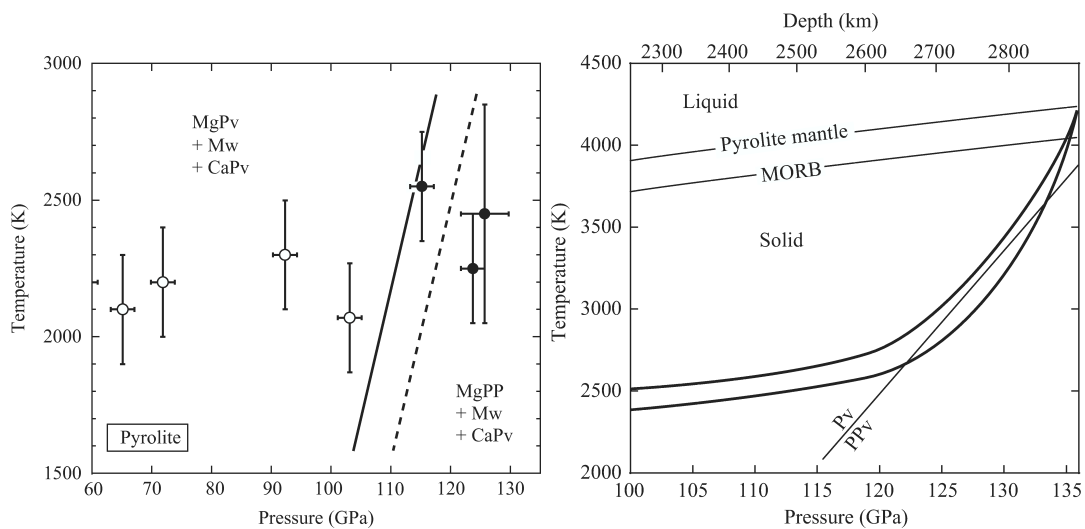
Left: phase relations near the bottom of the mantle for pyrolitic material (Hirose, 2007). The solid- and dashed line correspond to different pressure calibration of the HPT experiments. The Clapeyron slope of the phase boundary is assumed 11.5 MPa/K. CaPv: $CaSiO_3$ perovskite MgPv: $MgSiO_3$-rich perovskite, MgPP: $MgSiO_3$-rich post-perovskite, Mw: magnesiowüstite. Right: schematic temperature profiles in the lower mantle in relation to the perovskite (PV) to postperovskite (PPV) phase transition and the melting curve for pyrolitic mantle material and subducted basaltic crust (MORB) (Hirose *et al.* (2007) [606]).

This phase transition has a high valued positive slope of the phase boundary (Clapeyron parameter) $dP_t/dT \sim 10$ MPaK^{-1}. The intercept of the phase boundary with the core mantle boundary at ~ 136 GPa appears to be at a temperature several hundred Kelvin below the temperature of the liquid metal outer core as illustrated in the right part of the above figure. As a consequence the geotherm may intersect the phase boundary at multiple depth's, depending on the local mantle temperature, a phenomenon known as 'double crossing' (Hernlund *et al.* ., 2005). When a double crossing of the geotherm occurs, a thin layer exists directly bordering the core, where perovskite is the stable phase while on top of this bottom PV layer, a postperovskite layer exists with a variable thickness of up to several hundred kilometers.

A further implication of the phase diagram illustrated in is that the PPV layer will be absent in hot regions where the geotherm is completely above the PV-PPV phase boundary. This postperovskite phase boundary has also been associated with the top of the D" layer at variable height $\sim 100 - 300$ km above the CMB (Lay *et al.* ., 2005).

These seismological interpretations of the postperovskite phase boundary have been based on limited resolution methods applying 1-D radial velocity models. In a more recent development, techniques related to seismic wave migration methods, used in the oil and gas exploration industry,

are applied to delineate reflecting interfaces in 2-D and 3-D models in seismic stratigraphy of the CMB region (van der Hilst *et al.* ., 2007). This way the spatial resolution has been brought down to about 20 km, allowing mapping of detailed structures in the lowermost mantle. An important target of these high resolution seismic methods is the bottom interface of a postperovskite layer, associated with the 'double crossing', where mantle material transforms back from postperovskite into perovskite due to the steep increase in temperature in the bottom thermal boundary layer, illustrated in the figure above, related to the temperature contrast across the CMB.

In a similar way as for the spinel-postspinel phase transition the temperature at the seismic interfaces can then be estimated from the given depth(pressure) and the experimentally determined parameters of the postperovskite phase transition. This way a mantle adiabatic geotherm and boundary layer structure (error function) have been estimated with a CMB temperature $T_{cmb} \sim 4000$ K (van der Hilst *et al.* ., 2007). The 'foot' of the adiabatic mantle geotherm derived from this lies at a temperature of approximately 2500 K. Both the estimated CMB temperature and the foot of the adiabat seem to confirm independent earlier findings based on adiabatic temperature extrapolation over large depth ranges (Boehler, 1996 [113]).

The temperature contrast of about 1500 K across the core-mantle boundary resulting from these interpretations identify the bottom of the mantle as a thermal boundary layer, characteristic of a vigorously convecting layer where the boundary interface has a fixed or slowly varying temperature, as we will see in the section on heat transport in the mantle. As such these results from mineral physics and seismology have produced new evidence for strong mantle convective flow near the core-mantle boundary.

17.6 geostationary orbit
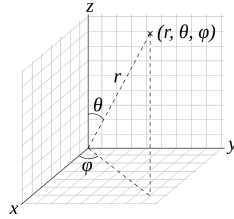
In spherical coordinates[10] the position, velocity and acceleration of a point are given by

$$
\begin{aligned}
\vec{r} &= r\vec{e}_r \\
\vec{v} &= \dot{r}\vec{e}_r + r\dot{\theta}\vec{e}_\theta + r\dot{\phi}\sin\theta\,\vec{e}_\phi \\
\vec{a} &= (\ddot{r} - r\dot{\theta}^2 - r\dot{\phi}^2\sin^2\theta)\vec{e}_r \\
&\quad + (r\ddot{\theta} + 2\dot{r}\dot{\theta} - r\dot{\phi}^2\sin\theta\cos\theta)\vec{e}_\theta \\
&\quad + (r\ddot{\phi}\sin\theta + 2\dot{r}\dot{\phi}\sin\theta + 2r\dot{\theta}\dot{\phi}\cos\theta)\vec{e}_\phi
\end{aligned}
$$

For an orbit at constant height ($r = R$, $\dot{r} = 0$), and constant angular velocities (i.e. $\ddot{\theta} = 0$ and $\ddot{\phi} = 0$) we arrive at

$$
\vec{v} = R\dot{\theta}\,\vec{e}_\theta + R\dot{\phi}\sin\theta\,\vec{e}_\phi \tag{17.71}
$$
$$
\vec{a} = (-R\dot{\theta}^2 - R\dot{\phi}^2\sin^2\theta)\,\vec{e}_r + (-R\dot{\phi}^2\sin\theta\cos\theta)\,\vec{e}_\theta + (2R\dot{\theta}\dot{\phi}\cos\theta)\,\vec{e}_\phi \tag{17.72}
$$

If the orbit is the equatorial plane, we have $\theta = \pi/2$, $\sin\theta = 1$, $\cos\theta = 0$ (and of course $\dot{\theta} = 0$) so now

$$
\vec{v} = R\dot{\phi}\,\vec{e}_\phi \tag{17.73}
$$
$$
\vec{a} = -R\dot{\phi}^2\,\vec{e}_r \tag{17.74}
$$

---

[10] https://en.wikipedia.org/wiki/Spherical_coordinate_system

The acceleration is the so-called centripetal[11] acceleration. We coin $\dot{\phi} = \omega$ is the constant angular velocity.

" A centripetal force (from Latin centrum, "center" and petere, "to seek") is a force that makes a body follow a curved path. The direction of the centripetal force is always orthogonal to the motion of the body and towards the fixed point of the instantaneous center of curvature of the path.

One common example involving centripetal force is the case in which a body moves with uniform speed along a circular path. The centripetal force is directed at right angles to the motion and also along the radius towards the centre of the circular path. The mathematical description was derived in 1659 by the Dutch physicist Christiaan Huygens.

In the case of an object that is swinging around on the end of a rope in a horizontal plane, the centripetal force on the object is supplied by the tension of the rope. The rope example is an example involving a 'pull' force.

Newton's idea of a centripetal force corresponds to what is nowadays referred to as a central force. When a satellite is in orbit around a planet, gravity is considered to be a centripetal force even though in the case of eccentric orbits, the gravitational force is directed towards the focus, and not towards the instantaneous center of curvature."[12]

" A geostationary orbit, also referred to as a geosynchronous equatorial orbit, is a circular geosynchronous orbit 35,786 km in altitude above Earth's equator, 42,164 km in radius from Earth's center, and following the direction of Earth's rotation. An object in such an orbit has an orbital period equal to Earth's rotational period, one sidereal day, and so to ground observers it appears motionless, in a fixed position in the sky. "[13]

The centripetal force of an orbiting body of mass m is then

$$\vec{F}_c = m\vec{a} = -mR\dot{\phi}^2\ \vec{e}_r$$

The gravitational force is

$$\vec{F}_g = -\frac{\mathcal{G}Mm}{R^2}\vec{e}_r$$

From Newton's second law of motion (sum of forces = mass * acceleration) we can write

$$m\vec{a} = -mR\dot{\phi}^2\ \vec{e}_r = -\frac{\mathcal{G}Mm}{R^2}\vec{e}_r$$

or,

$$R\dot{\phi}^2 = \frac{\mathcal{G}M}{R^2}$$

We have $\dot{\phi} = \omega = \frac{v}{R}$ so

$$R\left(\frac{v}{R}\right)^2 = \frac{\mathcal{G}M}{R^2}$$

$$v^2 = \frac{\mathcal{G}M}{R}$$

The velocity is given by $2\pi R/T$ where T is the desired period so now

$$\left(\frac{2\pi R}{T}\right)^2 = \frac{\mathcal{G}M}{R}$$

where T is the orbital period (i.e. one sidereal day), and is equal to 86164.09054 s. In the end

$$R = \left(\frac{\mathcal{G}MT^2}{4\pi^2}\right)^{1/3}$$

---

[11]Moving or tending to move towards a centre, as opposed to centrifugal: moving or tending to move away from a centre.

[12]https://en.wikipedia.org/wiki/Centripetal_force

[13]https://en.wikipedia.org/wiki/Geostationary_orbit

The resulting orbital radius is 42,164 kilometres. Subtracting the Earth's equatorial radius, 6,378 kilometres, gives the altitude of 35,786 kilometres.

The orbital speed is calculated by multiplying the angular speed by the orbital radius:

$$v = \omega R \simeq 3074.6 m/s$$

17.7 Programming exercises - February 2024

Background

We have seen that the calculation of the gravity vector and/or the gravity potential for a mass distribution in 3D space is of the form

$$\xi(\vec{r}) = \mathcal{G} \int_V f(\vec{r}, \vec{r}') \rho(\vec{r}') d\vec{r}'$$

where ξ is either g_x, g_y, g_z or U and f is a function of the coordinates $\vec{r}$ and $\vec{r}'$.

Let us now assume that the body under consideration can be subdivided into N_e smaller blocks/elements. By virtue of the linearity of the integral, we have

$$\xi(\vec{r}) = \mathcal{G} \sum_{e=1}^{N_e} \int_{V_e} f(\vec{r}, \vec{r}') \rho(\vec{r}') d\vec{r}'$$

We can further assume that inside each element the density is constant so that

$$\xi(\vec{r}) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e \int_{V_e} f(\vec{r}, \vec{r}') d\vec{r}'$$

We will now make a strong assumption which is only valid when elements are (very) small: we will assume that we can replace $f(\vec{r}, \vec{r}')$ by $f(\vec{r}, \vec{r}_e)$ where $\vec{r}_e$ is the location of the 'center' of the element. We then get:

$$\xi(\vec{r}) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e f(\vec{r}, \vec{r}_e) \int_{V_e} d\vec{r}'$$

And finally the integral term is simply the volume of the element V_e:

$$\xi(\vec{r}) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e f(\vec{r}, \vec{r}_e) V_e$$

In the end, assuming that the body of interest can be split into many small elements of constant density, the gravity fields at a location $\vec{r} = (x, y, z)$ can be computed as follows:

$$g_x(x, y, z) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e V_e \frac{x - x_e}{|\vec{r} - \vec{r}_e|^3} \tag{17.75}$$

$$g_y(x, y, z) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e V_e \frac{y - y_e}{|\vec{r} - \vec{r}_e|^3} \tag{17.76}$$

$$g_z(x, y, z) = \mathcal{G} \sum_{e=1}^{N_e} \rho_e V_e \frac{z - z_e}{|\vec{r} - \vec{r}_e|^3} \tag{17.77}$$

$$U(x, y, z) = -\mathcal{G} \sum_{e=1}^{N_e} \rho_e V_e \frac{1}{|\vec{r} - \vec{r}_e|} \tag{17.78}$$

where

$$|\vec{r} - \vec{r}_e| = \sqrt{(x - x_e)^2 + (y - y_e)^2 + (z - z_e)^2}$$

The following exercises are designed to test this approach which lends itself to numerical implementation. The basic idea is rather simple: generate a cloud of points in a regular manner such that we can assign them a corresponding volume and a density (and therefore a mass) when they are in the geometry of interest, and then use the formula above to compute the gravity vector and potential, and finally compare these values with the analytical solutions we derived for simple spherical bodies.

All quantities in the code(s) must be expressed in S.I. units, i.e. m, s, kg.

NO JUPYTER NOTEBOOK

Exercise 1: Full sphere

(1A) We consider a domain of size $2R \times 2R \times 2R$ centered on the origin. It is partitioned in $N \times N \times N$ cells as shown in the following figure.
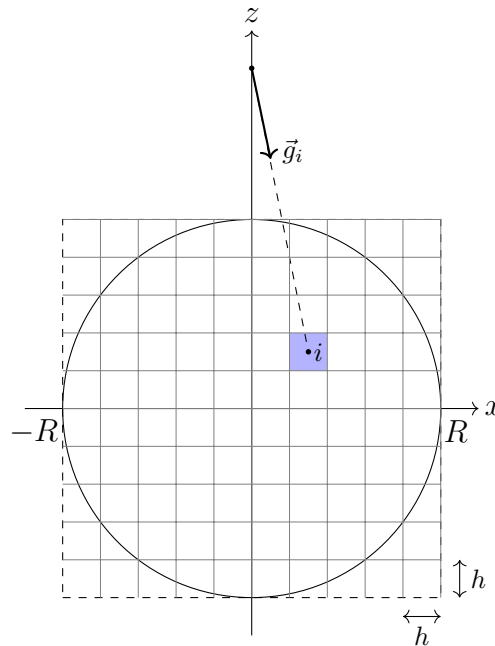
2D representation of the exercise. $N = 10$

Compute the total number of points `NP` as a function of `N`, the associated volume `dV` of a point (i.e. the volume of the cell the point is in) as a function of R and N and the size of a cell `h` as a function of R and N. Here is how your code should look like:

```
N=10
NP=
h=
dV=
```

(1B-1) To get started, start in 1D. Assume that we consider a 1D cube, i.e. the segment $[-R, R]$ that is divided into N cells. What is h? First, using the `linspace` function compute the x-coordinates of the N cell centers. Second, compute the same array `x` using a single for loop. This second approach will be the building block of the next question.

(1B-2) In the middle of each cell we place a point. Compute and store the coordinates of the N^3 points (use $R = 6371$ km). Please use arrays `x`, `y` and `z` to store the coordinates. how your code should look like:

```python
x = np.zeros(NP,dtype=np.float64) # x coordinates of all points
y = np.zeros(NP,dtype=np.float64) # y coordinates of all points
z = np.zeros(NP,dtype=np.float64) # z coordinates of all points
for i in range(?):
    for j in range(?):
        for k in range(?):
            x[?]=?
            y[?]=?
            z[?]=?
```

Important: these arrays are N^3 long because they contain the coordinates of *all* points (cell centers).

Help: draw on paper a $3 \times 3 \times 3$ elements grid. Place the axis system on the plot and explicitly write the arrays for all 27 points. Example:

```
x[0]=..... y[0]=..... z[0]=.....
x[1]=..... y[1]=..... z[1]=.....
x[2]=..... y[2]=..... z[2]=.....
...
x[26]=.... y[26]=.... z[26]=.....
```

Once you have done so, run your code for $N = 3$, print the arrays and compare their content with what you have on paper (tip: for this test temporarily set $R = 3$).

(1C) Assign a density $\rho_0 = 3000 \text{ kg m}^{-3}$ to points (cells) inside a sphere of radius R and zero otherwise, store these values in the `rho` array.
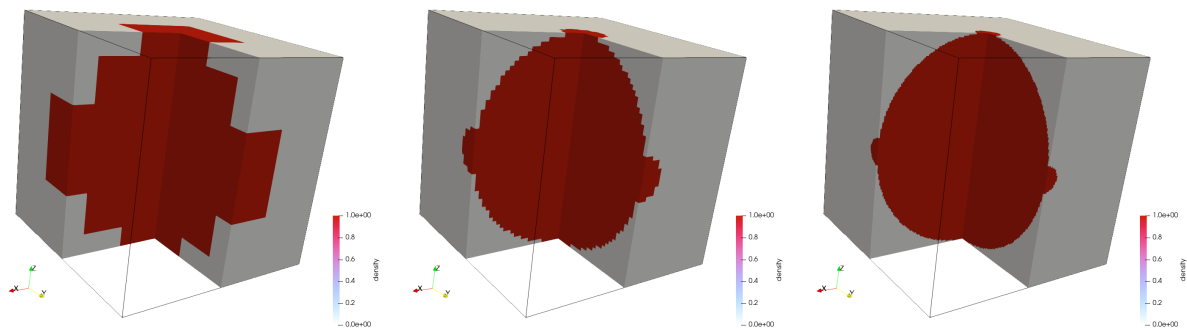
Example of density field ($\rho = 1$) for a 10^3, 50^3 and 150^3 mesh.

(1D) Fix $N = 10$. Compute the total mass of the sphere

$$M_s = \int_V \rho \, dV = \sum_e \rho_e V_e$$

and its volume[14]

$$V_s = \int_V dV = \sum_e V_e.$$

(1E) Fix $N = 10$. Compute the moment of inertia of the sphere with respect to rotation axis z using Eq. (17.1).

---

[14]Note that the sums run over the cells which center lies inside the sphere.

861

(1F) Repeat the last measurements (1D & 1E) with different values of $N \in (20, 30, 40, 50, ...?)$. For both the mass and moment of inertia report on the relative error as a function of h.

(1G) Compute the coordinates of 6 points situated at $z = 10^m$ meters *above the north pole* with $m = 0, 1, 2, 3, 4, 5$ and store these coordinates in arrays `xm`, `ym` `zm`.

```
xm = np.zeros(6,dtype=np.float64) # x coordinates of all points
ym = np.zeros(6,dtype=np.float64) # y coordinates of all points
zm = np.zeros(6,dtype=np.float64) # z coordinates of all points
```

(1H) Fix $N = 10$ for now. Compute the gravity potential U and acceleration vector components g_x, g_y, g_z at each of these 5 points using Eq. (17.14) (actually its discretised version, i.e. Eqs. (17.75), (17.76), (17.77) and (17.78).

(1I) Plot the computed quantities as a function of z and plot on the same graphic the analytical values.

(1J) Fix $m = 4$. Progressively increase N and record the absolute error on the gravity vector norm as a function of h. Plot this in log-log scale. Discuss.

(1K) Use the `prem_density` function to assign the PREM [380] density to the points. Compute the mass of the planet with this new density distribution and compare it with the mass of the Earth. Compute the gravity at the surface. hint: use a large(r) N for good results. How long are you willing to wait?

(1L) Bonus: time how long it takes to compute U, gx, gy, gz at a single location for $N = 20$. Report these times. Can you think of a way (and implement it) to arrive at the same results in less time?

Exercise 2: Hollow sphere

This is based on the previous exercise.

- For points with radius r such that $R/2 \leq r \leq R$ assign a density $\rho_0 = 3000 \text{kg m}^{-3}$ and zero otherwise.

- Compute the gravity potential and vector components on the x-axis between $r = 0$ and $r = 3R$ with steps of $R/100$.

- Plot the results and the analytical solution on the same plot as a function of r.

- Repeat the exercise with different values of N. Discuss.

Exercise 3: Full sphere - revisited

...NOT for 2024...

We are now going to re-do the first exercise but this time we do not want any point outside of the sphere. We shall therefore use the spherical coordinates (see Section 3.3.4). We will use three for loops, one over $r \in [0, R]$ values, one over $\theta \in [0, \pi]$ values and one over $\phi \in] - \pi, \pi]$ values. The number of points in each direction in this space is still N so that the total number of points is still N^3.

- Compute and store the coordinates of the points in the r, θ, ϕ space. Store these in arrays `r`, `theta`, `phi`.

- Use these coordinates to compute and store the Cartesian coordinates of these points.

- Plot this cloud of points in 3D. Discuss.

- Repeat the calculations of the first exercise.

- The cost of the calculation is the same as in exercise 1, but what about accuracy?

Report

The report should contain results from exercises 1 and 2. I expect one pdf file per student (maximum 10 pages) and the corresponding python file(s), all delivered in a single zip file. Your report does not have to follow questions 1A to 1K in sequential order. Please send all files in a single email per **March 29th, 2024, 23:59**.

You should have the following guidelines in mind when writing your report:

- Layout: is the document visually pleasing? Is it well structured? are the student names and numbers present ?

- Is there a complete bibliography (if/when applicable)?

- Introduction: is the context clear? Are the methods presented?

- Figures: Are they properly numbered? captioned? all figures must be referenced in the text. Are they of good quality? are they readable? are all axis labelled?

- Text: Overall quality of the language. Are there still typos ? Do all sentences make sense?

- If results are wrong, was an attempt made to document/explain the (probable) source of the problem?

- Discussion: are the results properly discussed, analyzed? are potential problems, errors, limitations discussed?

- Conclusion: Are the report's findings summarized and when applicable generalized?

For concrete examples of what not to do, check Appendix H.

17.8 Exam - February 2020

This is not the full 2020 exam, only the first two exercises.

Exercise 1

Let us consider a spherically symmetric body of radius R.

1. (1 pt) use symmetry considerations with regards to the 3 axis to arrive at the moment of inertia I:

$$I = \frac{8\pi}{3} \int_0^R \rho(r) r^4 dr \qquad (17.79)$$

2. (1/2 pt) The density inside the sphere is given by

$$\rho(r) = \rho_0 \left[a \left(\frac{r}{R} \right)^2 + b \left(\frac{r}{R} \right) + c \right]$$

Compute coefficients a, b, c such that the density matches the curve on the following figure:
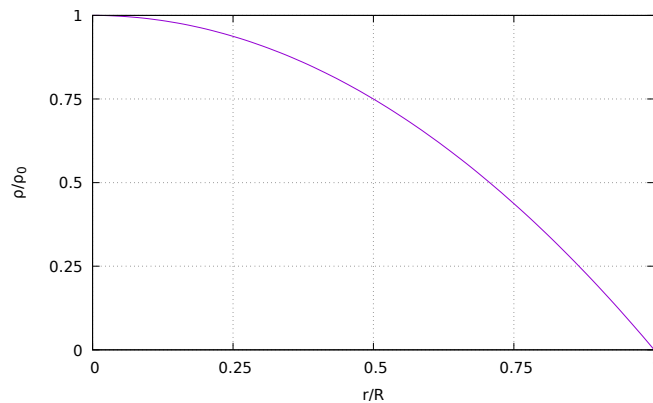
3. (1 pt) Compute the total mass M of the planet using this expression for the density.

4. (1 pt) Compute the moment of inertia I

5. (1/2 pt) Can I be written $I = fMR^2$? if so, give f.

6. (1/2 pt) What are the dimensions of ρ, I and M?

Exercise 1 - answer

Question 1 was treated in class. Looking at the figure, we see that $\rho(r = 0) = \rho_0$, i.e. $c = 1$. Also, we see that $\rho(r = R) = 0$, i.e.

$$\rho_0 \left[a + b + 1 \right] = 0$$

or, $a + b = -1$. Finally, we see that $\rho(r = R/2) = 3\rho_0/4$, so

$$\rho_0 \left[\frac{a}{4} + \frac{b}{2} + 1 \right] = \frac{3}{4}\rho_0$$

or, $a/4 + b/2 = -1/4$. This leads to $a = -1$ and $b = 0$ so

$$\rho(r) = \rho_0 \left[1 - \left(\frac{r}{R} \right)^2 \right]$$

The total mass is given by

$$
\begin{aligned}
M &= \int_V \rho(r) dV \\
&= 4\pi \int_0^R \rho(r) r^2 dr \\
&= 4\pi \int_0^R \rho_0 \left[1 - \left(\frac{r}{R}\right)^2 \right] r^2 dr \\
&= 4\pi \rho_0 \int_0^R \left[r^2 - \frac{r^4}{R^2} \right] dr \\
&= \frac{8\pi}{15} \rho_0 R^3 \tag{17.80}
\end{aligned}
$$

The moment of inertia I is

$$
\begin{aligned}
I &= \frac{8\pi}{3} \int_0^R \rho(r) r^4 dr \\
&= \frac{16\pi}{105} \rho_0 R^5 \tag{17.81}
\end{aligned}
$$

Finally,

$$
I = \frac{16\pi}{105} \rho_0 R^5 = \frac{2}{7} \left(\frac{8\pi}{15} \rho_0 R^3 \right) R^2
$$

so $f = 2/7$.

Exercise 2

Let us consider the same sphere as in the previous exercise, and the same density profile $\rho(r)$. The gravitational potential satisfies the Poisson equation:

$$
\Delta U = 4\pi \mathcal{G} \rho(\vec{r}) \tag{17.82}
$$

and we have the following relationship between the gravitational acceleration vector and the potential:
$\vec{g} = -\vec{\nabla} U$.

1. (1/2 pt) Write explicitely Eq.(17.82) for a point inside the sphere and a point outside the sphere.

2. (1 pt) Compute $g(r)$ and $U(r)$ for a point inside the sphere as a function of r. Use $\lim_{r \to 0} g(r) \neq \infty$ to get rid of an integration constant.

3. (1 pt) Compute $g(r)$ and $U(r)$ for a point outside the sphere as a function of r. Use $\lim_{r \to \infty} U(r) = 0$ to get rid of another integraion constant.

4. (1 pt) Use the continuity of $g(r)$ and $U(r)$ at $r = R$ to compute the last two remaining integration constants.

Exercise 2 - answer

Inside the sphere Eq.(17.82) is

$$
\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial U}{\partial r} \right) = 4\pi \mathcal{G} \rho_0 \left[a \left(\frac{r}{R} \right)^2 + b \left(\frac{r}{R} \right) + c \right]
$$

865

We could use the values of a, b and c obtained above but I here choose not to, in order to show that this exercise can be carried out independently from the first one. Then:

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 4\pi\mathcal{G}\rho_0\left[a\frac{r^4}{R^2} + b\frac{r^3}{R} + cr^2\right]$$

$$r^2\frac{\partial U}{\partial r} = 4\pi\mathcal{G}\rho_0\left[a\frac{r^5}{5R^2} + b\frac{r^4}{4R} + c\frac{r^3}{3}\right] + A$$

$$\frac{\partial U}{\partial r} = 4\pi\mathcal{G}\rho_0\left[a\frac{r^3}{5R^2} + b\frac{r^2}{4R} + c\frac{r}{3}\right] + \frac{A}{r^2}$$

so that

$$g(r) = -\frac{\partial U}{\partial r} = -4\pi\mathcal{G}\rho_0\left[a\frac{r^3}{5R^2} + b\frac{r^2}{4R} + c\frac{r}{3}\right] + \frac{A}{r^2}$$

We use $\lim_{r\to 0} g(r) \neq \infty$ to arrive at $A = 0$. Finally

$$U_{in}(r) = 4\pi\mathcal{G}\rho_0\left[a\frac{r^4}{20R^2} + b\frac{r^3}{12R} + c\frac{r^2}{6}\right] + B$$

With $a = -1$, $b = 0$ and $c = 1$:

$$g_{in}(r) = -4\pi\mathcal{G}\rho_0\left[-\frac{r^3}{5R^2} + \frac{r}{3}\right]$$

$$U_{in}(r) = 4\pi\mathcal{G}\rho_0\left[-\frac{r^4}{20R^2} + \frac{r^2}{6}\right] + B$$

Outside the sphere we must solve

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 0$$

$$r^2\frac{\partial U}{\partial r} = C$$

$$\frac{\partial U}{\partial r} = \frac{C}{r^2}$$

$$U_{out}(r) = -\frac{C}{r} + D$$

We use $\lim_{r\to\infty} U(r) = 0$ to arrive at $D = 0$ so that

$$g_{out}(r) = -\frac{C}{r^2}$$

$$U_{out}(r) = -\frac{C}{r}$$

Both fields should match at $r = R$:

$$g_{in}(r = R) = g_{out}(r = R)$$
$$U_{in}(r = R) = U_{out}(r = R)$$

i.e.

$$-4\pi\mathcal{G}\rho_0\left[a\frac{R^3}{5R^2} + b\frac{R^2}{4R} + c\frac{R}{3}\right] = -\frac{C}{R^2}$$

so

$$C = 4\pi\mathcal{G}\rho_0 R^3 \left[\frac{a}{5} + \frac{b}{4} + \frac{c}{3}\right]$$

Now,

$$4\pi\mathcal{G}\rho_0 \left[a\frac{R^4}{20R^2} + b\frac{R^3}{12R} + c\frac{R^2}{6}\right] + B = -\frac{C}{R}$$

$$4\pi\mathcal{G}\rho_0 R^3 \left[\frac{a}{20} + \frac{b}{12} + \frac{c}{6}\right] + BR = -4\pi\mathcal{G}\rho_0 R^3 \left[\frac{a}{5} + \frac{b}{4} + \frac{c}{3}\right]$$

so

$$BR = -4\pi\mathcal{G}\rho_0 R^3 \left[\frac{a}{5} + \frac{a}{20} + \frac{b}{4} + \frac{b}{12} + \frac{c}{3} + \frac{c}{6}\right]$$

$$B = -4\pi\mathcal{G}\rho_0 R^2 \left[\frac{a}{4} + \frac{b}{3} + \frac{c}{2}\right]$$

With $a = -1$, $b = 0$ and $c = 1$:

$$B = -\pi\mathcal{G}\rho_0 R^2 \qquad C = \frac{8\pi}{15}\mathcal{G}\rho_0 R^3$$

Finally

$$
\begin{aligned}
g_{in}(r) &= -4\pi\mathcal{G}\rho_0 \left[-\frac{r^3}{5R^2} + \frac{r}{3}\right] \\
U_{in}(r) &= 4\pi\mathcal{G}\rho_0 \left[-\frac{r^4}{20R^2} + \frac{r^2}{6}\right] - \pi\mathcal{G}\rho_0 R^2 \\
g_{out}(r) &= -\frac{8\pi}{15}\mathcal{G}\rho_0 R^3 \frac{1}{r^2} \\
U_{out}(r) &= -\frac{8\pi}{15}\mathcal{G}\rho_0 R^3 \frac{1}{r}
\end{aligned}
$$

Also, remembering that the mass of the planet is $M = \frac{8\pi}{15}\rho_0 R^3$ so that

$$
\begin{aligned}
g_{out}(r) &= -\frac{\mathcal{G}M}{r^2} \\
U_{out}(r) &= -\frac{\mathcal{G}M}{r}
\end{aligned}
$$

No surprise there...

17.9   Exam - March 2021

Exercise 1

Let us consider a spherically symmetric body of radius R.

1. (1 pt) use symmetry considerations with regards to the 3 axis to arrive at the moment of inertia I:

$$I = \frac{8\pi}{3} \int_0^R \rho(r) r^4 dr \qquad (17.83)$$

2. (1 pt) The density inside the sphere is given by

$$\rho(r) = a\frac{r}{R} + b$$

   Compute the total mass M of the planet using this expression for the density.

3. (1 pt) Compute the moment of inertia I also as a function of a and b.

4. (1/2 pt) Can I be written $I = fMR^2$? if so, give f.

5. (1/2 pt) What are the dimensions of a, b, ρ, I and M?

6. (1/2 pt) Set $a = 0$ and $b = \rho_0$ and look again at M and I. Conclude.

Exercise 1 - answer

The first question is answered in Problem 1, see Section 17.3.

   The mass of the planet is given by

$$
\begin{aligned}
M &= \iiint_V \rho(r) dV \\
&= \iiint_V \rho(r) r^2 \sin\theta \, dr d\theta d\phi \\
&= \iiint_V \left(a\frac{r}{R} + b \right) r^2 \sin\theta \, dr d\theta d\phi \\
&= 4\pi \int_0^R \left(a\frac{r}{R} + b \right) r^2 dr \\
&= 4\pi \left[\frac{a}{R} \int_0^R r^3 dr + b \int_0^R r^2 dr \right] \\
&= 4\pi \left[\frac{a}{R}\frac{1}{4}R^4 + b\frac{1}{3}R^3 \right] \\
&= 4\pi R^3 \left(\frac{a}{4} + \frac{b}{3} \right) \qquad (17.84)
\end{aligned}
$$

The moment of inertia of the planet is given by

$$
\begin{aligned}
I &= \frac{8\pi}{3} \iiint_0^R \rho(r) r^4 dV \\
&= \frac{8\pi}{3} \iiint_0^R \left(a\frac{r}{R} + b \right) r^4 dV \\
&= \frac{8\pi}{3} \left(\frac{a}{6} + \frac{b}{5} \right) R^5 \\
&= \underbrace{(4\pi R^3) \left(\frac{a}{4} + \frac{b}{3} \right)}_{M} \left(\frac{a}{4} + \frac{b}{3} \right)^{-1} R^2 \frac{2}{3} \left(\frac{a}{6} + \frac{b}{5} \right) \\
&= \underbrace{\left(\frac{a}{4} + \frac{b}{3} \right)^{-1} \frac{2}{3} \left(\frac{a}{6} + \frac{b}{5} \right)}_{f} MR^2 \qquad\qquad (17.85)
\end{aligned}
$$

The dimensions of a, b, I and M are as follows:

$$
[a] = [b] = ML^{-3} \qquad [I] = ML^2 \qquad [M] = M
$$

When $a = 0$ and $b = \rho_0$ we find the standard mass of a constant density sphere $M = \frac{4}{3}\pi R^3 \rho_0$ and $f = 2/5$.

Exercise 2

Let us consider the same sphere as in the previous exercise, and the same density profile $\rho(r)$. The gravitational potential satisfies the Poisson equation:

$$
\Delta U = 4\pi \mathcal{G} \rho(\vec{r}) \qquad\qquad (17.86)
$$

and we have the following relationship between the gravitational acceleration vector and the potential: $\vec{g} = -\vec{\nabla} U$.

1. (1/2 pt) Write explicitely Eq.(17.86) for a point inside the sphere and a point outside the sphere.

2. (1 pt) Compute $g(r)$ and $U(r)$ for a point inside the sphere as a function of r. Use $\lim_{r\to 0} g(r) \neq \infty$ to get rid of an integration constant.

3. (1 pt) Compute $g(r)$ and $U(r)$ for a point outside the sphere as a function of r. Use $\lim_{r\to\infty} U(r) = 0$ to get rid of another integraion constant.

4. (1 pt) Use the continuity of $g(r)$ and $U(r)$ at $r = R$ to compute the last remaining integration constants.

5. (1/2 pt) Set $a = 0$ and $b = \rho_0$ and sketch the obtained fields.

Exercise 2 - answer

outside $\Delta U = 0$. Inside $\Delta U = 4\pi\mathcal{G}(a\frac{r}{R} + b)$

$$
\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial U}{\partial r} \right) = 4\pi\mathcal{G} \left(a\frac{r}{R} + b \right)
$$

869

$$\Rightarrow \qquad \frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 4\pi\mathcal{G}\left(a\frac{r^3}{R} + br^2\right)$$

$$\Rightarrow \qquad r^2\frac{\partial U}{\partial r} = 4\pi\mathcal{G}\left(a\frac{r^4}{4R} + b\frac{r^3}{3}\right) + A$$

$$\Rightarrow \qquad \frac{\partial U}{\partial r} = 4\pi\mathcal{G}\left(a\frac{r^2}{4R} + b\frac{r}{3}\right) + \frac{A}{r^2}$$

so that

$$g(r) = -\frac{\partial U}{\partial r} = -4\pi\mathcal{G}\left[a\frac{r^2}{4R} + b\frac{r}{3}\right] + \frac{A}{r^2}$$

We use $\lim_{r\to 0} g(r) \neq \infty$ to arrive at $A = 0$. Finally

$$g_{in}(r) = -4\pi\mathcal{G}\left(a\frac{r^2}{4R} + b\frac{r}{3}\right) \qquad (17.87)$$

and

$$U_{in} = -\int g(r)dr = 4\pi\mathcal{G}\left(a\frac{r^3}{12R} + b\frac{r^2}{6}\right) + B$$

Outside the sphere we must solve

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 0$$

$$\Rightarrow \qquad r^2\frac{\partial U}{\partial r} = C$$

$$\Rightarrow \qquad \frac{\partial U}{\partial r} = \frac{C}{r^2}$$

$$\Rightarrow \qquad U_{out}(r) = -\frac{C}{r} + D$$

We use $\lim_{r\to\infty} U(r) = 0$ to arrive at $D = 0$ so that

$$g_{out}(r) = -\frac{C}{r^2}$$
$$U_{out}(r) = -\frac{C}{r}$$

Both fields should match at $r = R$:

$$g_{in}(r = R) = g_{out}(r = R)$$
$$U_{in}(r = R) = U_{out}(r = R)$$

i.e.

$$-4\pi\mathcal{G}\left(a\frac{R^2}{4R} + b\frac{R}{3}\right) = -\frac{C}{R^2} \qquad (17.88)$$

$$\Rightarrow \qquad C = 4\pi\mathcal{G}R^3\left(\frac{a}{4} + \frac{b}{3}\right) = M\mathcal{G}$$

so unsurprisingly (!):

$$g_{out}(r) = -\frac{M\mathcal{G}}{r^2}$$
$$U_{out}(r) = -\frac{M\mathcal{G}}{r}$$

and

$$4\pi\mathcal{G}\left(a\frac{R^3}{12R}+b\frac{R^2}{6}\right)+B=-\frac{M\mathcal{G}}{R}$$

$$4\pi\mathcal{G}R^2\left(a\frac{1}{12}+b\frac{1}{6}\right)+B=-\frac{M\mathcal{G}}{R}$$

$$B=-\frac{M\mathcal{G}}{R}-4\pi\mathcal{G}R^2\left(a\frac{1}{12}+b\frac{1}{6}\right)$$

so

$$U_{in}=4\pi\mathcal{G}\left(a\frac{r^3}{12R}+b\frac{r^2}{6}\right)-\frac{M\mathcal{G}}{R}-4\pi\mathcal{G}R^2\left(a\frac{1}{12}+b\frac{1}{6}\right)$$

If $a=0$ and $b=\rho_0$, then

$$g_{in}(r)=-4\pi\mathcal{G}\rho_0\frac{r}{3}$$

and

$$U_{in}=4\pi\mathcal{G}\left(\rho_0\frac{r^2}{6}\right)-\frac{M\mathcal{G}}{R}-4\pi\mathcal{G}R^2\left(\rho_0\frac{1}{6}\right)=\frac{2\pi}{3}\mathcal{G}\rho_0 r^2-\frac{3}{2}\frac{M\mathcal{G}}{R}$$

which is the solution to problem 14.

17.10   Gravity benchmarks

There are many analytical solutions for buried bodies of simple shape. Hereafter are the most common ones:

17.10.1   Buried sphere (3D)

To calculate the pull of gravity, we can use the fact that a sphere has the same gravitational pull as a point mass located at its centre. The distance between the measurement point and the center of the sphere is $\sqrt{x^2+d^2}$, so

$$g_z=\frac{\mathcal{G}M_{sphere}d}{(x^2+d^2)^{3/2}}$$

Let us take the following example: radius a=50m, $\Delta\rho=2000$, variable depth d=100m
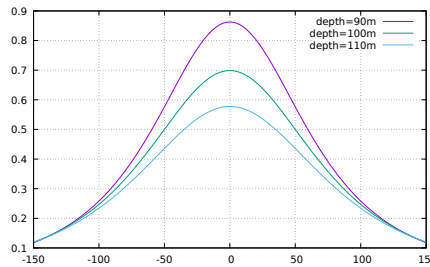
g_z has its maximum value directly above the sphere at $x=0$m and is given by

$$g_z^{max}=\frac{\mathcal{G}M_{sphere}d}{(d^2)^{3/2}}=\frac{\mathcal{G}M_{sphere}}{d^2}$$

We can then find the half width of the curve by finding $x_{1/2}$ such that

$$\frac{\mathcal{G}M_{sphere}d}{(x_{1/2}^2+d^2)^{3/2}}=\frac{g_z^{max}}{2}=\frac{\mathcal{G}M_{sphere}}{2d^2}$$

or, FINISH , derive $x_{1/2}$

17.10.2 Buried horizontal cylinder (3D)

anticline can be approximated by a horizontal cylinder

$$g_z = \frac{2\mathcal{G}\pi a^2 d\Delta\rho}{x^2 + d^2}$$

the maximum value of g z is located directly above the axis of the cylinder

g zmax for a cylinder is larger than g zmax for a sphere of the same radius.

Cannot distinguish a buried sphere from a cylinder with just a single profile. Need to collect gravity on a grid and make a map.

17.10.3 Buried column (2D)

$$g_z = 2\mathcal{G}\Delta\rho b \ln \frac{r_2}{r_1}$$
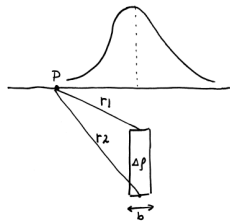
17.10.4 Buried columns (2D)

$$g_z = 2\mathcal{G}\sum_i \Delta\rho_i b_i \ln \frac{r_{2,i}}{r_{1,i}}$$

in ./images/gravity/columns/
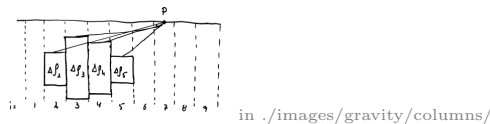
17.10.5 Uniform layer of rock

A layer of rock has an infinite extent, thickness Δz and a density ρ. The gravitational attraction of this slab at the point P at height z above the layer is

$$g_z = 2\pi\mathcal{G}\rho\Delta z$$

Note that g z does not depend on the distance from the layer to the measurement point.

17.10.6 A constant density shell (Root *et al.* , 2021)

Results & raw data in `./images/benchmark_gravity/bench1`

The shell is defined between R_1 and R_2. It contains a single material of density 3300kg m^{-3}. The layer is centered around depth 100km. Gravity is measured 250km above the surface, i.e. $r = 6621$km. The thickness of the shell is 2, 5 or 10km.

The analytical gravity vector norm is given by

$$g = \frac{4\pi}{3}\rho\frac{R_2^3 - R_1^3}{r^2}\mathcal{G}$$

where we take $\mathcal{G} = 6.67384 \cdot 10^{-11}$ (default in ASPECT) or $\tilde{\mathcal{G}} = 6.67428 \cdot 10^{-11}$ sometimes.

shell thickness (km)	volume (m^3)	mass (kg)	gravity using $\mathcal{G}$ (m s^{-2})	gravity using $\tilde{\mathcal{G}}$ (m s^{-2})
2	9.8835614e17	3.2615753e21	496.542034795	496.574771345
5	2.4708905e17	8.1539385e21	1241.35514223	1241.43698361
10	4.9417817e18	1.630788e22	2482.71067903	2482.87436182

In the ASPECT input file there are three main parameters which may influence the results:

- the radial resolution, controlled in the input file by: `set Number of slices = 1,2,3,4`

- the tangential/lateral resolution, controlled by: `set Initial lateral refinement = 3,4,5,6`

- the number of (additional) quadrature points, controlled by: `set Quadrature degree increase =0,1,...6`

We set here the default values at 1, 6 and 3 respectively.

	lat. res. 3	lat. res. 4	lat. res. 5	lat. res. 6	lat. res. 7
nslice=1 (1 cells radial) # cells	384	1,536	6,144	24,576	98,304
	6×64	6×256	$6 \times 1,024$	$6 \times 4,096$	$6 \times 16,384$
	6×8^2	6×16^2	6×32^2	6×64^2	6×128^2
nslice=2 (2 cells radial) # cells	768	3,072	12,288	49,152	196,608
nslice=3 (3 cells radial) # cells	1,152	4,608	18,432	73,728	294,912
nslice=4 (4 cells radial) # cells	1,536	6,144	24,576	98,304	393,216
average area (m2)	1.328292e+12	3.320732e11	8.30183e10	2.075457e10	5.188644e9
approx size (km)	1152km	576	288km	144km	72km
approx size (degree)	10.5	5.2	2.6	1.3	0.65

Earth has a surface of $\mathcal{S} = 4\pi R^2 \simeq 5.1006447 \cdot 10^{14}\text{m}^2$. An average degree resolution means that this surface would be tesselated in blocks of approximately $2\pi R/360 \simeq 111$km size. There would then be about 41,398 of such blocks. If a resolution of 2 degrees is required, then the blocks would be about 220km in size and there would be about 10,349 blocks.

Results obtained with ASPECT with $\tilde{G}$ are in the following table:

Thickness (km)	2	5	10
Shell formula (mGal)	496.574771345	1241.43698361	2482.87436182
$m = 4, \sim 5°$	496.554320/496.602897/496.574854	1241.385829/1241.507337/1241.437190	2482.771870/2483.015344/2482.874775
$m = 5, \sim 2.6°$	496.574748/496.574819/496.574771	1241.436926/1241.437102/1241.436983	2482.874246/2482.874599/2482.874361
$m = 6, \sim 1.3°$	496.574771/496.574771/496.574771	1241.436984/1241.436984/1241.436984	2482.874362/2482.874362/2482.874362

Results for a 10km thick shell with Aspect
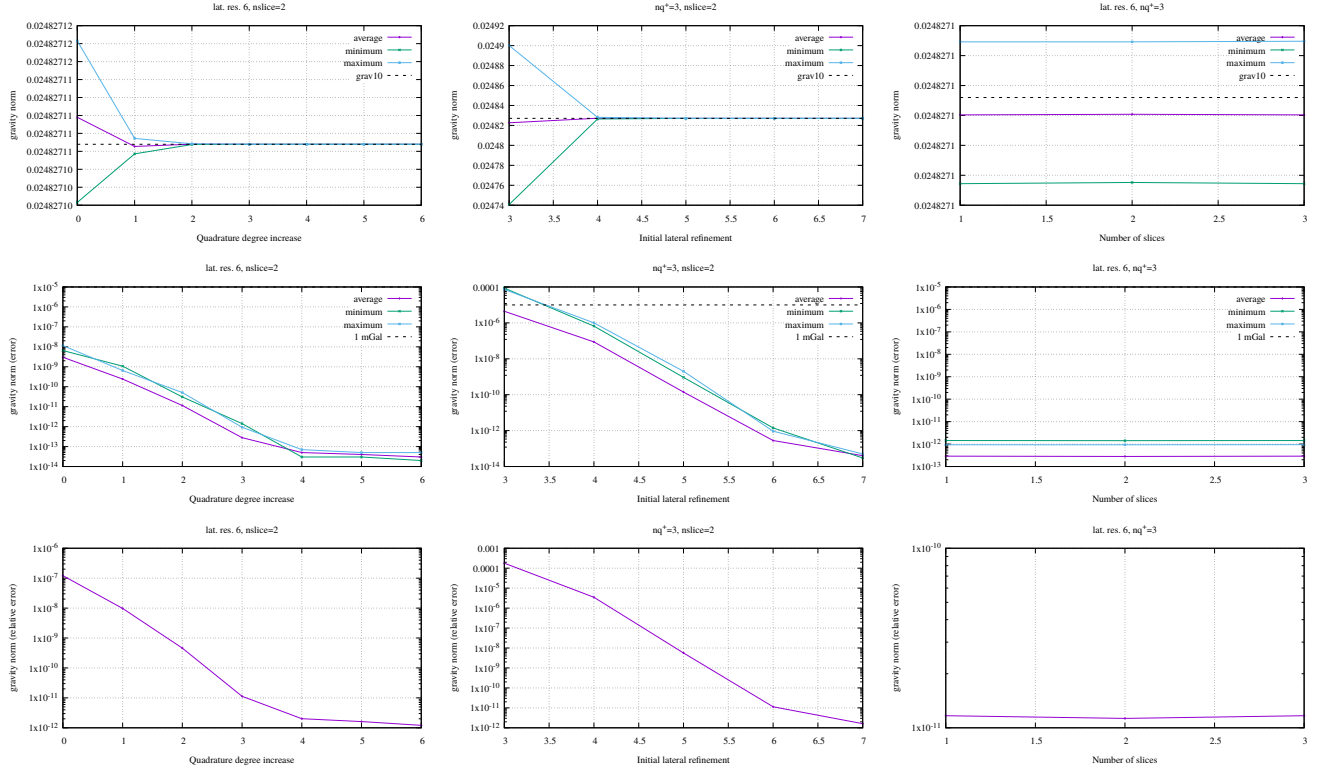
in ./images/benchmark_gravity/bench1/

I then define the concept of 'cost'. In terms of computational cost, there is a tradeoff between resolution and number of quadrature points. The cost is then defined as

$$C = nel \times nq^3$$

where nel is the number of elements in the mesh and nq is the number of quadrature points per element and per dimension.
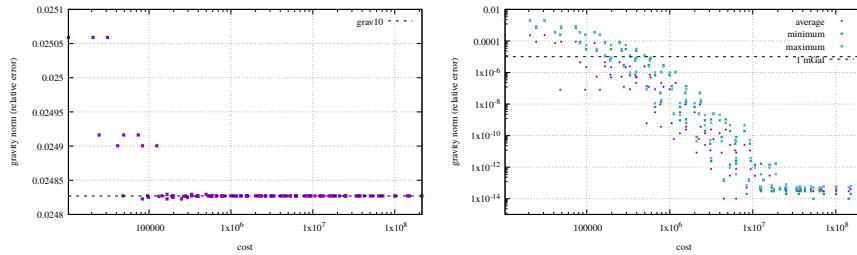
in ./images/benchmark_gravity/bench1/

Preliminary conclusion: nq+ at least 3, nslice not so important here, lat res at least 6
TODO RUN 2 and 5 km shells !

17.10.7 The WINTERC mono-layer benchmark (Root *et al.*, 2021)

A single data file is used, `rho_56km_SH_W32.txt`, stored in `images/benchmark_gravity/bench2`. It contains density values for a single layer comprised between 56 and 80km depths, i.e. there is no radial variation of the density. Because of how Aspect works, the density values need to be transformed into initial temperatures. Using the simple material model we have

$$\rho = \rho_0(1 - \alpha(T - T_0))$$

so

$$T = \frac{1}{\alpha}\left(1 - \frac{\rho}{\rho_0}\right)$$

and we here take $\alpha = 3 \cdot 10^{-5}$, $T_0 = 0$ and $\rho_0 = 3300$, so that the temperatures range between -1198.9 and 141.5. These values make no sense, but all we want is that the densities $\rho(T)$ generated by the material model are those of the original dataset.

Furthermore, the `rho_56km_SH_W32.txt` file contains 720x360 lines, i.e. the resolution is a half degree for longitude and latitude. These range from 0.25 to 359.75 and from -89.75 to 89.75 respectively. These must be transformed into spherical coordinates $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$.

Also the original data file contains longitude, latitude and density values for a thick layer. The ascii data file which Aspect can read requires radial values in increasing order as well, so for each combination $\phi - \theta$ I generate two values, one at radius 6371-81km and one at 6371-55km so that the depth layer 56-80km fits in it. The data format of the ascii file is specified in `data/initial-temperature/ascii-data/test/shell_3d.txt` in Aspect.

Stone 98 reads in `rho_56km_SH_W32.txt` and generates the `bench2.txt` file which is to be read by Aspect. Note that the first line of this file is mandatory and reads: `# POINTS: 2 720 360`

in ./images/benchmark_gravity/bench2/
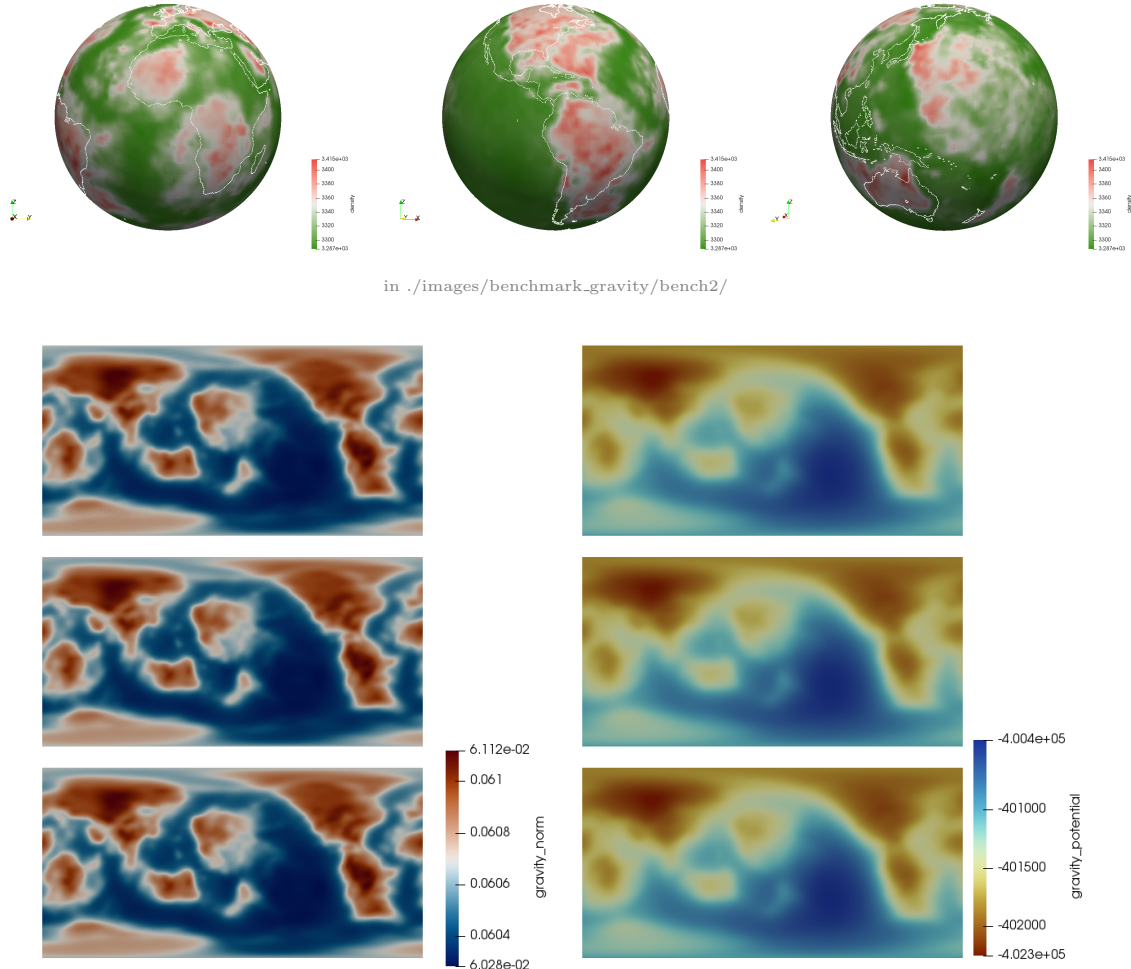
Results obtained with ASPECT . Top to bottom, level 4, 5, 6. Measurements grid is 181x91 points.
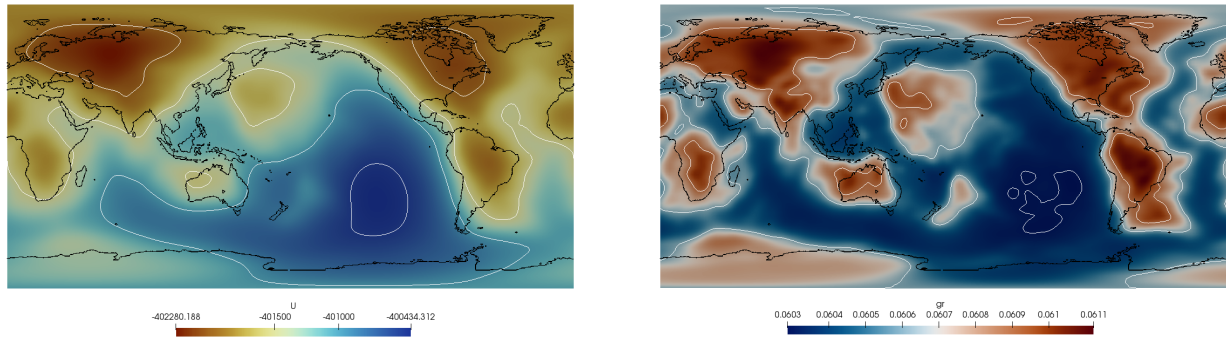
From Stone 98. Resolution of measurement grid is 181x91. It took about 19,100 seconds to run, averaging 1.16s per measurement point. Potential isocontours at -400.5e3, -401e3, -401.5e3, -402e3. Radial acceleration contours at 0.0603, 0.0606 and 0.0609.

level	avrg density	total mass	number of cells	time
4	3323	3.981e+22	1,536	503s
5	3323	3.981e+22	6,144	2030s
6	3323	3.981e+22	24.576	8190s

17.10.8 Moho benchmark (Root *et al.* , 2021)

We consider an 80km thick shell with a density interface inside, using CRUST1.0 Moho for the boundary (upper dens = 2900kg/m3, lower dens = 3300 kg/m3). Stone 97 reads the CRUST1.0 file and transforms it into `bench3.ascii` in the right ASPECT format.
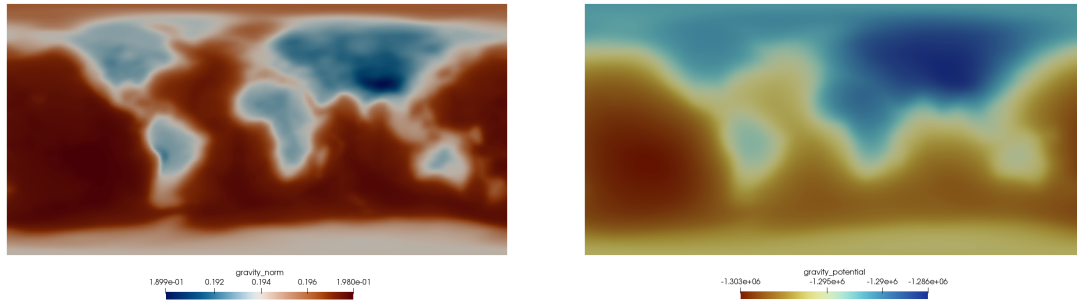
bench3.ascii file has 301 points in the radial direction, i.e. a 300m resolution. 181x91 gravity measurement points. Lateral refinement level is 6, 25 slices.

in ./images/benchmark_gravity/bench3/

17.10.9 Gravity potential and gravity field of a two-layer spherically symmetric planet

Note: see also problem 16 in previous section.

Let us assume that the planet consists of two layers: the core (of density ρ_c) and the mantle (of density ρ_m). Its outer radius is R_2 and the cmb is at R_1. We wish to compute the gravitational potential of this system and we start from the spherically symmetric Poisson equation:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 4\pi\mathcal{G}\rho \qquad (17.89)$$

We have three domains on the r-axis (a stands for air):

- $r \in [0, R_1]$, density ρ_c

- $r \in [R_1, R_2]$, density ρ_m

- $r \in [R_2, +\infty)$, density $\rho_a = 0$

We denote $\hat{\rho} = 4\pi\mathcal{G}\rho$. In the end we have to solve Eq. 17.89 in all three domains:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = \hat{\rho}_c \quad r \in [0, R_1] \qquad (17.90)$$

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = \hat{\rho}_m \quad r \in [R_1, R_2] \qquad (17.91)$$

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 0 \quad r \in [R_2, +\infty)] \qquad (17.92)$$

The generic solution of Eq. (17.89) is obtained as follows:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = \hat{\rho} \qquad (17.93)$$

$$\Rightarrow \quad \frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = \hat{\rho}r^2 \qquad (17.94)$$

$$\Rightarrow \quad r^2\frac{\partial U}{\partial r} = \frac{1}{3}\hat{\rho}r^3 + A \qquad (17.95)$$

$$\Rightarrow \quad \frac{\partial U}{\partial r} = \frac{1}{3}\hat{\rho}r + \frac{A}{r^2} = g(r) \qquad (17.96)$$

$$\Rightarrow \quad U(r) = \frac{1}{6}\hat{\rho}r^2 - \frac{A}{r} + B \qquad (17.97)$$

So now we have the following set of equations:

$$g(r) = \frac{1}{3}\hat{\rho}_c r + \frac{A}{r^2} \qquad U(r) = \frac{1}{6}\hat{\rho}_c r^2 - \frac{A}{r} + B \qquad r \in [0, R_1] \qquad (17.98)$$

$$g(r) = \frac{1}{3}\hat{\rho}_m r + \frac{C}{r^2} \qquad U(r) = \frac{1}{6}\hat{\rho}_m r^2 - \frac{C}{r} + D \qquad r \in [R_1, R_2] \qquad (17.99)$$

$$g(r) = \frac{1}{3}\hat{\rho}_a r + \frac{E}{r^2} \qquad U(r) = \frac{1}{6}\hat{\rho}_a r^2 - \frac{E}{r} + F \qquad r \in [R_2, +\infty) \qquad (17.100)$$

We know that $\rho_a = 0$ and we additionally impose

$$\lim_{r\to 0} g(r) = 0 \qquad \lim_{r\to+\infty} U(r) = 0$$

which automatically leads to $A = F = 0$.

In the end we have:

$$g(r) = \frac{1}{3}\hat{\rho}_c r \qquad\qquad U(r) = \frac{1}{6}\hat{\rho}_c r^2 + B \qquad\qquad r \in [0, R_1] \qquad (17.101)$$

$$g(r) = \frac{1}{3}\hat{\rho}_m r + \frac{C}{r^2} \qquad\qquad U(r) = \frac{1}{6}\hat{\rho}_m r^2 - \frac{C}{r} + D \qquad\qquad r \in [R_1, R_2] \qquad (17.102)$$

$$g(r) = \frac{E}{r^2} \qquad\qquad U(r) = -\frac{E}{r} \qquad\qquad r \in [R_2, +\infty) \qquad (17.103)$$

We have now four unknowns B, C, D, E. Since the two fields g and U must be continuous at $r = R_1$ and $r = R_2$, then we have four constraints and that will allow us to determine the four unknowns.

Let us start with the continuity at $r = R_1$:

$$g(r = R_1) \;\; = \;\; \frac{1}{3}\hat{\rho}_c R_1 = \frac{1}{3}\hat{\rho}_m R_1 + \frac{C}{R_1^2} \qquad (17.104)$$

$$U(r = R_1) \;\; = \;\; \frac{1}{6}\hat{\rho}_c R_1^2 + B = \frac{1}{6}\hat{\rho}_m R_1^2 - \frac{C}{R_1} + D \qquad (17.105)$$

The first equation yields

$$C = \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m)$$

which we plug in the second equation:

$$\frac{1}{6}\hat{\rho}_c R_1^2 + B = \frac{1}{6}\hat{\rho}_m R_1^2 - \frac{R_1^2}{3}(\hat{\rho}_c - \hat{\rho}_m) + D$$

$$\frac{1}{6}(\hat{\rho}_c - \hat{\rho}_m)R_1^2 + B = -\frac{R_1^2}{3}(\hat{\rho}_c - \hat{\rho}_m) + D$$

$$(\frac{1}{6} + \frac{1}{3})(\hat{\rho}_c - \hat{\rho}_m)R_1^2 + B = D$$

$$\frac{1}{2}(\hat{\rho}_c - \hat{\rho}_m)R_1^2 + B = D$$

We cannot go any further so we turn to $r = R_2$:

$$g(r = R_2) \;\; = \;\; \frac{1}{3}\hat{\rho}_m R_2 + \frac{C}{R_2^2} = \frac{E}{R_2^2} \qquad (17.106)$$

$$U(r = R_2) \;\; = \;\; \frac{1}{6}\hat{\rho}_m R_2^2 - \frac{C}{R_2} + D = -\frac{E}{R_2} \qquad (17.107)$$

in which we insert the known value of C:

$$\frac{1}{3}\hat{\rho}_m R_2 + \frac{1}{R_2^2}\frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m) = \frac{E}{R_2^2} \qquad (17.108)$$

$$\frac{1}{6}\hat{\rho}_m R_2^2 - \frac{1}{R_2}\frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m) + D = -\frac{E}{R_2} \qquad (17.109)$$

Multiplying the first equation by R_2^2 and the second one by R_2:

$$\frac{1}{3}\hat{\rho}_m R_2^3 + \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m) = E \qquad (17.110)$$

$$\frac{1}{6}\hat{\rho}_m R_2^3 - \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m) + DR_2 = -E \qquad (17.111)$$

The first equation then gives us E:

$$E = \frac{1}{3}\hat{\rho}_m(R_2^3 - R_1^3) + \frac{1}{3}R_1^3\hat{\rho}_c$$

So in the end, we have

$$C = \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m) \tag{17.112}$$

$$E = \frac{1}{3}\hat{\rho}_m(R_2^3 - R_1^3) + \frac{1}{3}R_1^3\hat{\rho}_c \tag{17.113}$$

$$D = \frac{1}{R_2}\left(-E - \frac{1}{6}\hat{\rho}_m R_2^3 + \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m)\right) \tag{17.114}$$

$$= \frac{1}{R_2}\left(-\frac{1}{3}\hat{\rho}_m(R_2^3 - R_1^3) - \frac{1}{3}R_1^3\hat{\rho}_c - \frac{1}{6}\hat{\rho}_m R_2^3 + \frac{R_1^3}{3}(\hat{\rho}_c - \hat{\rho}_m)\right) \tag{17.115}$$

$$= \frac{1}{6R_2}\left[(-2R_2^3 + 2R_1^3 - R_2^3 - 2R_1^3)\hat{\rho}_m + (-2R_1^3 + 2R_1^3)\hat{\rho}_c\right] \tag{17.116}$$

$$= \frac{1}{6R_2}\left(-3R_2^3\hat{\rho}_m\right) \tag{17.117}$$

$$= -\frac{R_2^2}{2}\hat{\rho}_m \tag{17.118}$$

$$B = D - \frac{1}{2}(\hat{\rho}_c - \hat{\rho}_m)R_1^2 \tag{17.119}$$

$$= -\frac{R_2^2}{2}\hat{\rho}_m - \frac{1}{2}(\hat{\rho}_c - \hat{\rho}_m)R_1^2 \tag{17.120}$$

$$= \frac{1}{2}(R_1^2 - R_2^2)\hat{\rho}_m - \frac{1}{2}R_1^2\hat{\rho}_c \tag{17.121}$$

We set $R_1 = 3400$km, $R_2 = 6400km$, $\rho_c = 6000$kg/m3 and $\rho_m = 4000$kg/m3 and obtain the following fields:
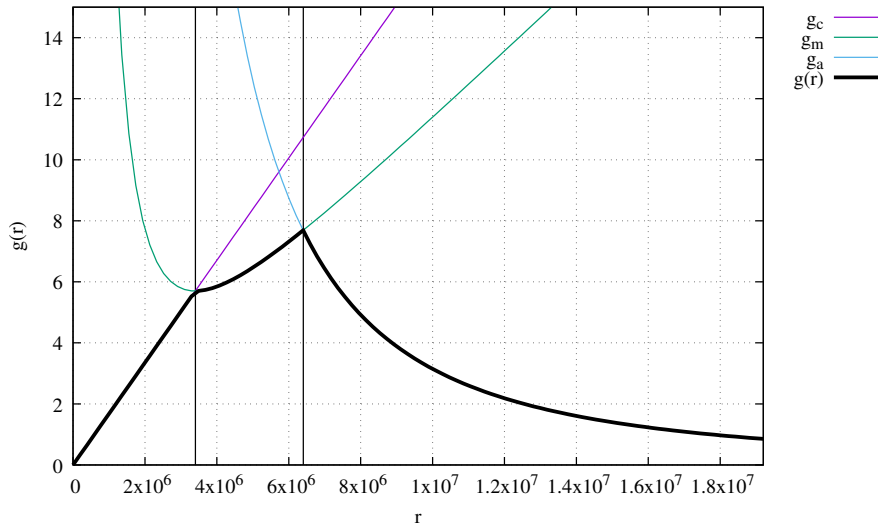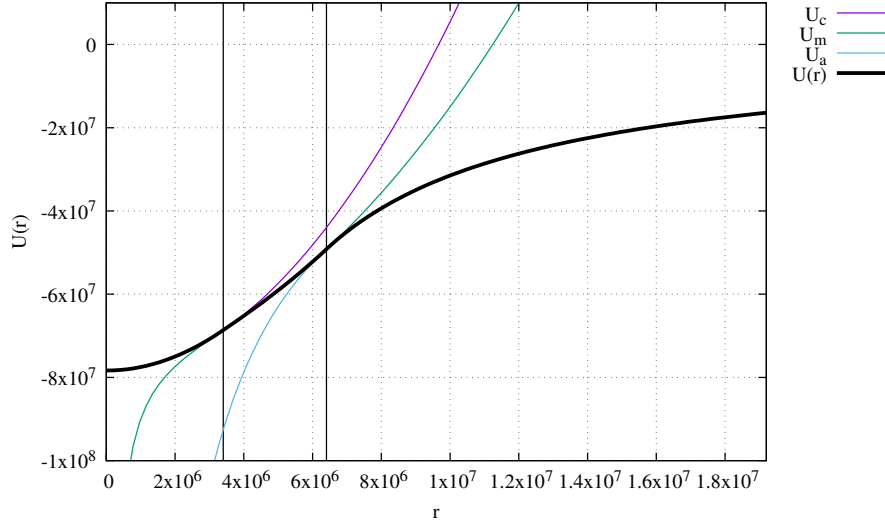
We find that the fields fulfill all conditions and are continuous, as expected.

Also, setting $\rho_c = 0$ (i.e. the planet is a hollow sphere), we recover the following fields:
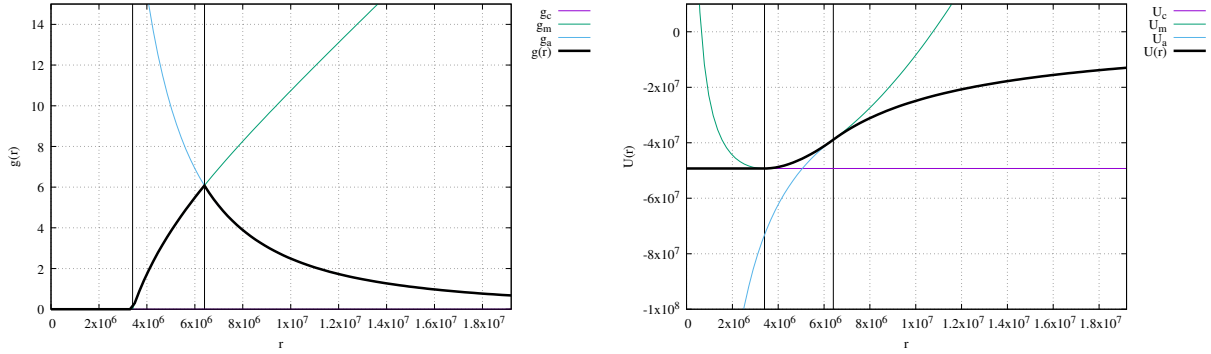
These are similar to the results presented in Appendix A of Thieulot [1344].

We can also consider the case of a constant density planet, i.e. $\rho_m = \rho_c = \rho_0 = 5000\text{kg/m3}$ In that case $C = 0$,

$$E = \frac{1}{3}\hat{\rho}_0 R_2^3 = \frac{4\pi}{3}\mathcal{G}\rho_0 R_2^3 = M\mathcal{G}$$

$$D = -\frac{1}{2}R_2^2\hat{\rho}_0$$

$$B = -\frac{1}{2}R_2^2\hat{\rho}_0$$

so that

$$g(r) = \frac{4\pi}{3}\mathcal{G}\rho_0 r \qquad U(r) = \frac{1}{6}\hat{\rho}_0 r^2 - \frac{1}{2}R_2 2\hat{\rho}_0 \qquad r \in [0, R_1] \tag{17.122}$$

$$g(r) = \frac{4\pi}{3}\mathcal{G}\rho_0 r \qquad U(r) = \frac{1}{6}\hat{\rho}_0 r^2 - \frac{1}{2}R_2^2\hat{\rho}_0 \qquad r \in [R_1, R_2] \tag{17.123}$$

$$g(r) = \frac{\mathcal{G}M}{r^2} \qquad U(r) = -\frac{\mathcal{G}M}{r} \qquad r \in [R_2, +\infty) \tag{17.124}$$

Unsurprisingly we obtain the same expressions inside the core and the mantle and we recover the expressions of problem 14 (see previous section).

In the end the fields are as follows:
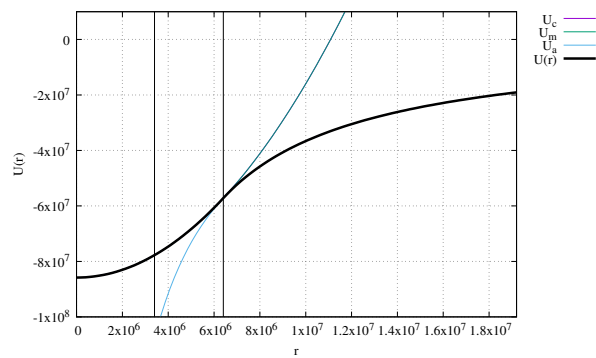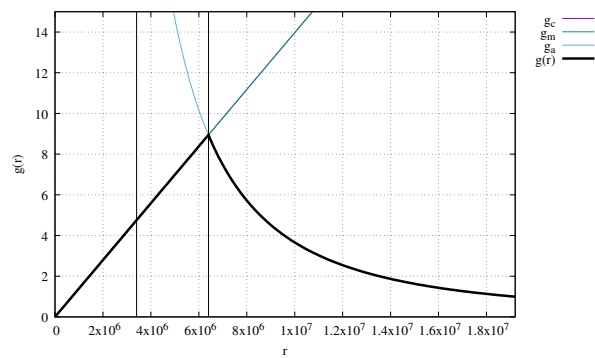
17.11 Gravity forward calculations in practice

$$\vec{g}(\vec{r}) = -\vec{\nabla} U(\vec{r})$$

with

$$U(\vec{r}) = \mathcal{G} \int \int \int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} dV$$

with (in a Cartesian coordinates system):

$$|\vec{r} - \vec{r}'| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2} \qquad dV = dx' dy' dz'$$

We then have

$$
\begin{aligned}
g_x(x, y, z) &= -\mathcal{G} \int \int \int \frac{\partial}{\partial x} \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} dx' dy' dz' \\
&= \mathcal{G} \int \int \int \frac{\rho(\vec{r}')(x - x')}{|\vec{r} - \vec{r}'|^3} dx' dy' dz' \\
g_y(x, y, z) &= -\mathcal{G} \int \int \int \frac{\partial}{\partial y} \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} dx' dy' dz' \\
&= \mathcal{G} \int \int \int \frac{\rho(\vec{r}')(y - y')}{|\vec{r} - \vec{r}'|^3} dx' dy' dz' \\
g_z(x, y, z) &= -\mathcal{G} \int \int \int \frac{\partial}{\partial z} \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} dx' dy' dz' \\
&= \mathcal{G} \int \int \int \frac{\rho(\vec{r}')(z - z')}{|\vec{r} - \vec{r}'|^3} dx' dy' dz'
\end{aligned}
$$

In order to compute the gravity tensor $\boldsymbol{T} = \vec{\nabla} \vec{g}$, also called gravitational gravity tensor [1164] or Marussi tensor, we need to compute $\frac{\partial^2 U}{\partial \alpha \partial \beta}$ with $\alpha, \beta = x, y, z$ and we obtain (See Arroyo *et al.* (2015) [31]):

$$T_{xx}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(x - x')^2 - (\vec{r} - \vec{r}')^2}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.125)$$

$$T_{yy}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(y - y')^2 - (\vec{r} - \vec{r}')^2}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.126)$$

$$T_{zz}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(z - z')^2 - (\vec{r} - \vec{r}')^2}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.127)$$

$$T_{xy}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(x - x')(y - y')}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.128)$$

$$T_{xz}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(x - x')(z - z')}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.129)$$

$$T_{yz}(x, y, z) = -\mathcal{G} \int \int \int \frac{3(y - y')(z - z')}{|\vec{r} - \vec{r}'|^5} \rho(\vec{r}') dx' dy' dz' \qquad (17.130)$$

Note that the trace satisfies the Laplace equation $T_{xx} + T_{yy} + T_{zz} = \Delta U = 0$.

redo all calculations to be sure

Unless the geometry is conveniently chosen with lots of symmetry and the density is also very simple the above integrals cannot be computed analyticaly and one must resort to numerical integration based on a tesselation of the space, such as prisms or tesseroids.

Prisms

The gravitational potential U of a right rectangular parallelopiped (prism) of homogeneous mass-density ρ_0 is described by Newton's integral [590]

$$U(x,y,z) = \mathcal{G}\rho_0 \int_{z_1}^{z_2} \int_{y_1}^{y_2} \int_{x_1}^{x_2} \frac{1}{|\vec{r} - \vec{r'}|} dx' dy' dz' = \mathcal{G}\rho_0 \int_{z_1}^{z_2} \int_{y_1}^{y_2} \int_{x_1}^{x_2} \frac{1}{\sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}} dx' dy' dz'$$
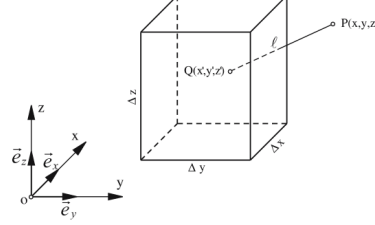
Taken from [590].

The denominator is the distance between the computation point $P(x, y, z)$ and the running integration point $Q(x', y', z')$. The coordinate axes have been assumed to be parallel to the edges of the prism, which extends between the coordinate surfaces related to the bounds x_1, x_2, y_1, y_2, z_1, z_2. It is well known that the integral can be solved analytically (Mader (1951) [868], see also Nagy et al. [976, 975]), resulting in the formula for the potential $U(x, y, z)$:

$$U(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} (-1)^{i+j+k} \left(A + B + C - \frac{1}{2}D \right)$$

with

$$A = (x - x_i)(y - y_j) \ln \left| \frac{z - z_k + r_{ijk}}{\sqrt{(x - x_i)^2 + (y - y_j)^2}} \right|$$

$$B = (y - y_j)(z - z_j) \ln \left| \frac{x - x_i + r_{ijk}}{\sqrt{(y - y_j)^2 + (z - z_k)^2}} \right|$$

$$C = (x - x_i)(z - z_k) \ln \left| \frac{y - y_j + r_{ijk}}{\sqrt{(z - z_k)^2 + (x - x_i)^2}} \right|$$

$$D = (x - x_i)^2 \arctan \frac{(y - y_j)(z - z_k)}{(x - x_i)r_{ijk}}$$

$$+ (y - y_j)^2 \arctan \frac{(z - z_k)(x - x_i)}{(y - y_j)r_{ijk}}$$

$$+ (z - z_k)^2 \arctan \frac{(x - x_i)(y - y_j)}{(z - z_k)r_{ijk}}$$

and

$$r_{ijk} = \sqrt{(x - x_i)^2 + (y - y_j)^2 + (z - z_k)^2}$$

Remark. *The direct application of this equation will fail when the computation point P is situated on an edge or on a corner of the prism; the respective limit values have been derived by Nagy et al. [976, 975]*

The gravity vector and tensor can then be computed [1069, 31, 298]:

$$g_x(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k} \times$$

$$\left[(y-y_j)\ln((z-z_k)+r_{ijk}) + (z-z_k)\ln((y-y_j)+r_{ijk}) - (x-x_i)\arctan\frac{(y-y_j((z-z_k))}{(x-x_i)r_{ijk}}\right]$$

$$g_y(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k} \times$$

$$\left[(z-z_k)\ln((x-x_i)+r_{ijk}) + (x-x_i)\ln((z-z_k)+r_{ijk}) - (y-y_j)\arctan\frac{(x-x_i)(z-z_k)}{(y-y_j)r_{ijk}}\right]$$

$$g_z(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k} \times$$

$$\left[(y-y_j)\ln((x-x_i)+r_{ijk}) + (x-x_i)\ln((y-y_j)+r_{ijk}) - (z-z_k)\arctan\frac{(x-x_i)(y-y_j)}{(z-z_k)r_{ijk}}\right]$$

$$T_{xx}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\arctan\frac{(y-y_j)(z-z_k)}{(x-x_i)r_{ijk}}$$

$$T_{yy}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\arctan\frac{(x-x_i)(z-z_k)}{(y-y_j)r_{ijk}}$$

$$T_{zz}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\arctan\frac{(x-x_i)(y-y_j)}{(z-z_k)r_{ijk}}$$

$$T_{xy}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\ln((z-z_k)+r_{ijk})$$

$$T_{xz}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\ln((y-y_j)+r_{ijk})$$

$$T_{yz}(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\ln((x-x_i)+r_{ijk})$$

Note that Heck & Seitz [590] report that the logarithmic terms can be transformed in order to provide a better numerical stability and then

$$g_x(x,y,z) = \mathcal{G}\rho_0 \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left[(y-y_j)\ln\left|\frac{(z-z_k)+r_{ijk}}{\sqrt{(x-x_i)^2+(y-y_j)^2}}\right|\right.$$

$$\left.+(z-z_k)\ln\left|\frac{(y-y_j)+r_{ijk}}{\sqrt{(x-x_i)^2+(z-z_k)^2}}\right| - (x-x_i)\arctan\frac{(y-y_j)(z-z_k)}{(x-x_i)r_{ijk}}\right] \quad (17.131)$$

The gravitational potential of the homogeneous rectangular prism, neglecting terms of order four and higher in x, y, z, is then given by MacMillan's (1930)[15] formula:

$$U(x,y,z) = \mathcal{G}\rho_0\Delta_x\Delta_y\Delta_y\left[\frac{1}{l_0} + \frac{3(x_0-x)^2-l_0^2}{24l_0^5}\Delta_x^2 + \frac{3(y-y_0)^2-l_0^2}{24l_0^5}\Delta_y^2 + \frac{3(z-z_0)^2-l_0^2}{24l_0^5}\Delta_z^2 + \mathcal{O}(\Delta^4)\right]$$

[15]MacMillan WD (1930) Theoretical Mechanics, vol 2: the The- ory of the potential. McGraw-Hill, New York (reprinted by Dover Publications, New York 1958)
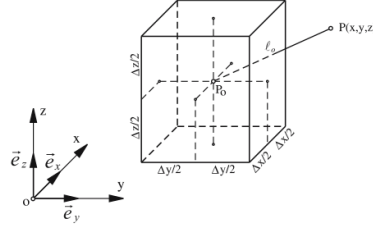
Taken from [590].

It is obvious that the zero-order approximation is identical with the potential of a point-mass at P_0 when the total mass of the prism $m = \rho_0 \Delta_x \Delta_y \Delta_z$ is concentrated at its geometrical centre P_0:

$$U(x, y, z) = \mathcal{G} \rho_0 \Delta_x \Delta_y \Delta_y \frac{1}{l_0}$$

It is also common [371] to look at:

- the differential curvature magnitude (DCM) which is also known as the horizontal directive tendency, computed by a combination of components of tensor T_{xx}, T_{xy} and T_{yy}. It emphasizes greatly the effects of shallower sources (Saad, 2006);

$$DCM = \sqrt{(T_{xx} - T_{yy})^2 + 2T_{xy}^2}$$

- the horizontal gradient magnitude (HGM) of g_z can be computed from the horizontal derivative components of g_z and can be used as edge detector or to map the body outline as it verifies the prism boundaries

$$HGM = \sqrt{T_{zx}^2 + T_{zy}^2} = \sqrt{\left(-\frac{\partial g_z}{\partial x}\right)^2 + \left(-\frac{\partial g_z}{\partial y}\right)^2}$$

- the total gradient magnitude (TGM) is computed from the three derivatives of vertical component of gravity:

$$TGM = \sqrt{T_{zx}^2 + T_{zy}^2 + T_{zz}^2} = \sqrt{\left(-\frac{\partial g_z}{\partial x}\right)^2 + \left(-\frac{\partial g_z}{\partial y}\right)^2 + \left(-\frac{\partial g_z}{\partial z}\right)^2}$$

Example 1 The result of calculating the components of a prism measuring 200m^3 at a height of 0.01 km, with an observation mesh of 1km×1km, and discretized every 20m is shown hereunder:
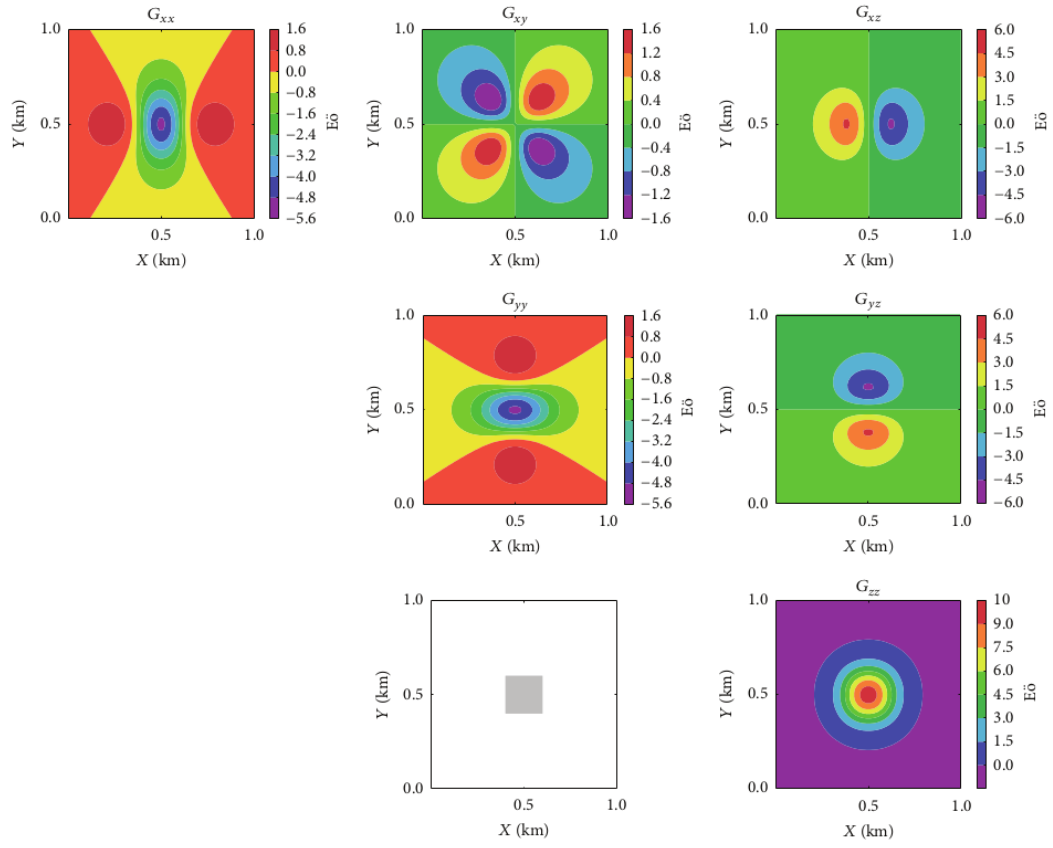
Taken from Arroyo *et al.* (2015) [31]. Gravity gradient response for a prism buried a depth of 100m,

Each side having a length of 200 m and constant density contrast of 100 kg/m^3.

Example 2 . Buried prims of size 8x4x1 km along x, y and z directions respectively. Density is 2700. Mapped gravity field and its gradient on a plane of constant 1000m height.
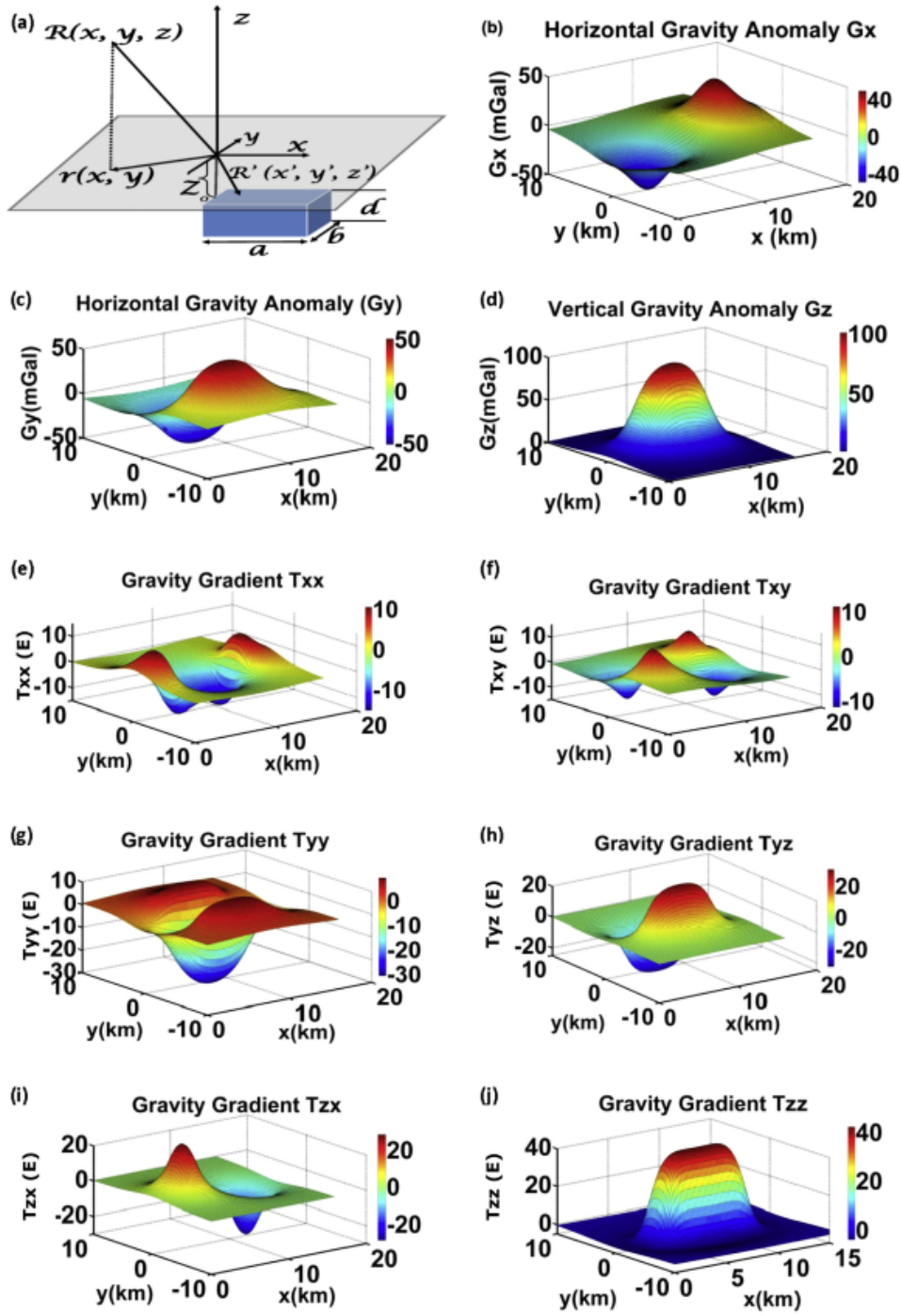
(a) A model containing a prism and (b-d): corresponding gravity vector components and (e-j) GGT components with sampling interval of 0.2 km in x and y directions. Taken from [371]
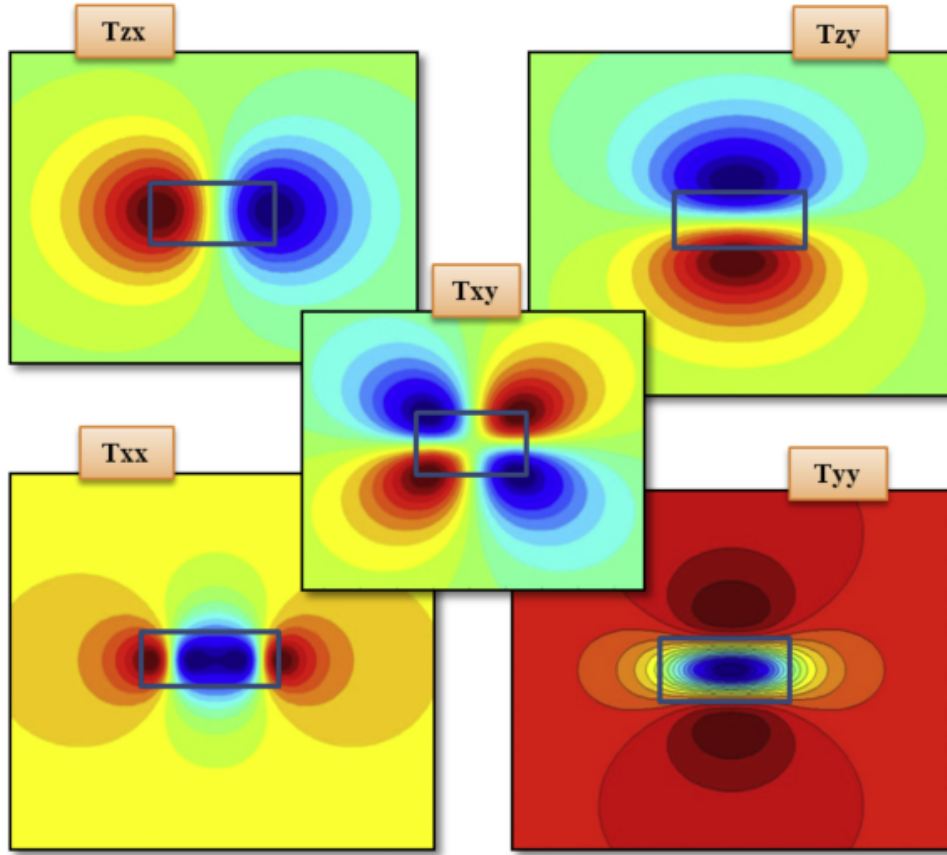
A map view of complex behavior of gravity gradients for prism model. Taken from [371]
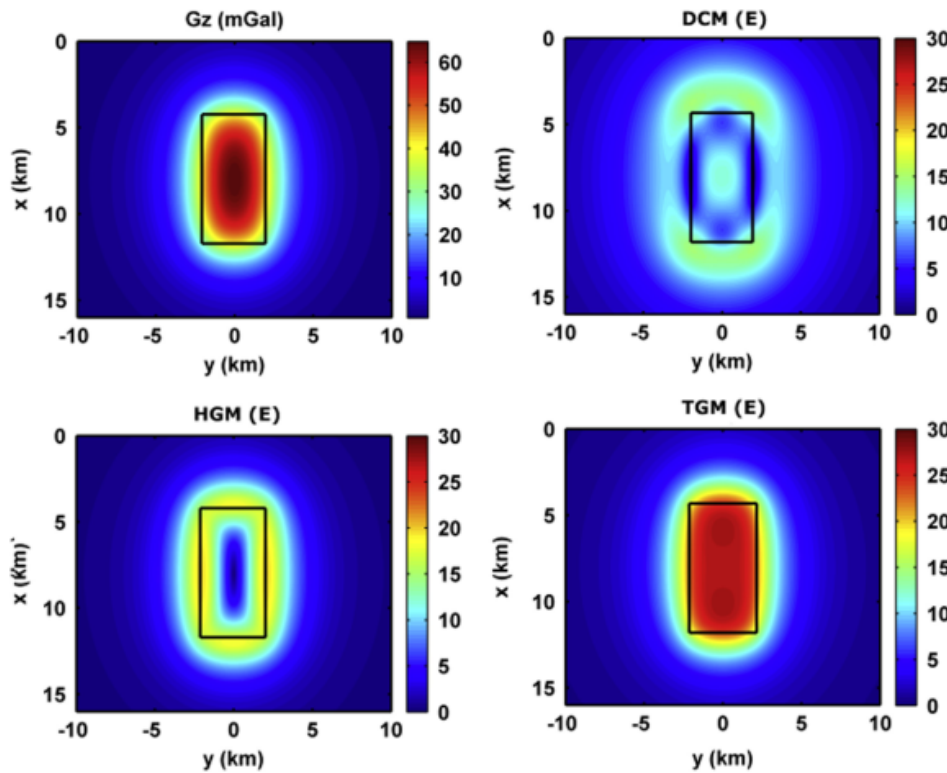
Computed vertical gravity component G z and three invariants map of HGM, DCM and TGM for given prism model. HGM = Horizontal Gradient Magnitude, DCM = Differential Curvature Magnitude and TGM = Total Gradient Magnitude. Taken from [371]

📚 Relevant Literature:

- Analytic Expressions for the Gravity Gradient Tensor of 3D Prisms with Depth-Dependent Density [684]

889

- New computationally efficient quadrature formulas for triangular prism elements [778]

- 3D Gravity Modeling of Complex Salt Features in the Southern Gulf of Mexico [986].

- Spherical prism gravity effects by Gauss-Legendre quadrature integration [33]

- Perturbing effects of sub-lithospheric mass anomalies in GOCE gravity gradient and other gravity data modelling: Application to the Atlantic-Mediterranean transition zone [450]

Tesseroids

A tesseroid, or spherical prism, is segment of a sphere.
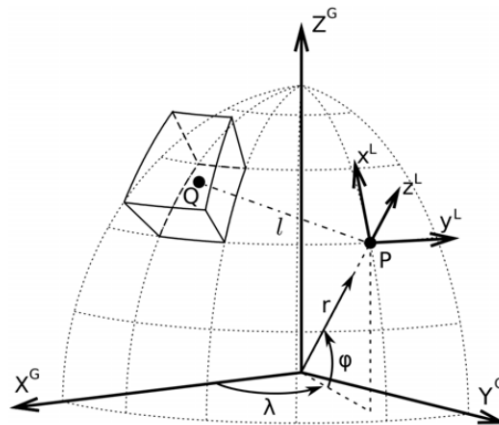
Taken from [1388]

Tesseroids is a collection of command-line programs for modeling the gravitational potential, acceleration, and gradient tensor. Tesseroids supports models and computation grids in Cartesian and spherical coordinates. `https://tesseroids.readthedocs.io/en/stable/`

📦 Relevant Literature:

- Forward modeling and inversion of gravitational fields in spherical coordinates, L. Uieda, Phd Thesis [1387]

- Tesseroids: Forward-modeling gravitational fields in spherical coordinates [1386]

- Optimal forward calculation method of the Marussi tensor due to a geologic structure at GOCE height [1388]

- Optimized formulas for the gravitational field of a tesseroid [527]

Tetrahedra

Werner and Scheeres [1450] (1997) derived analytical expressions for the gravity potential, field and tensor generated by any polyhedron. These can be applied to tetrahedra (See STONE 113).

Metherell and Quinn [915] (1986) derived analytical expressions for the gravity field generated by so-called 111-tetrahedra. Note the corrections in Carré, Metherell, and Quinn [225] (1986).

Chanut, Aljbaae, and Carruba [235] (2015) used a mascon approach (point mass approach?) on tetrahedra making up an asteroid:
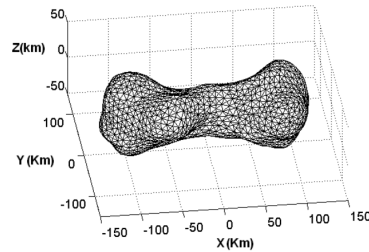
Taken from Chanut, Aljbaae, and Carruba [235]. Polyhedron model 3D of asteroid (216) Kleopatra. The shape was built with 4092 faces.

📚 Relevant Literature:

Manik Talwani, J Lamar Worzel, and Mark Landisman. "Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone". In: *J. Geophys. Res.: Solid Earth* 64.1 (1959), pp. 49–59

M. Talwani and M. Ewing. "Rapid Computation of Gravitational Attraction of Three-Dimensional Bodies of Arbitrary Shape". In: *Geophysics* 25.1 (1960), pp. 203–225

17.12   Instruments to measure gravity

Gravity meters

include here pic of Vening Meinesz
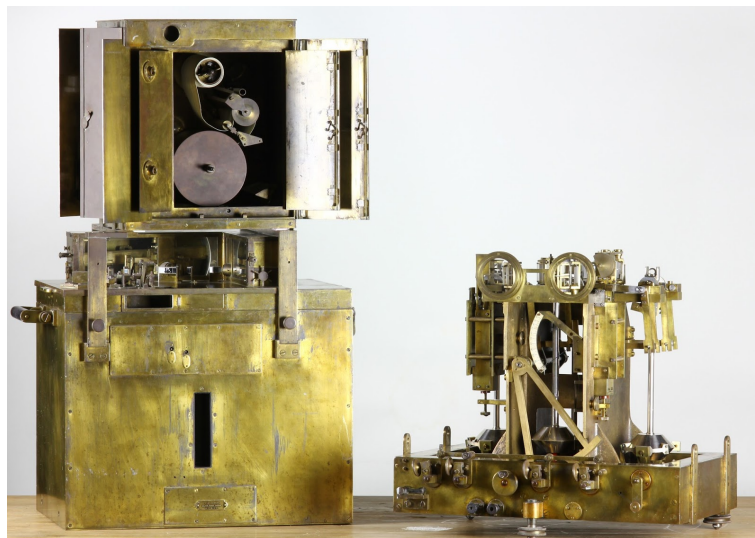
Taken from website[16]. The pendulum apparatus of Vening Meinesz, also known as "Het Gouden Kalf" (the Golden Calf). Positioned on the left side is the protective casing with the recording instrument on top. On the right side is the pendulum apparatus with the three pendulums at the back.

See video by Bart Root: `https://youtu.be/SVTJA3KAnck?si=-OZ1lPnHQwHyOkEl`

Absolute gravity measurements   After a time t an object has fallen by a distance x in a gravity field g with $x = gt^2/2$ so that $g = 2x/t^2$.

---

[16]`http://deepearthscience.blogspot.com/2016/06/the-gravimeter-of-professor-vening.html`

by Micro-g LaCoste. The FG5[17] operates by using a free-fall method. An object is dropped inside a vacuum chamber and its position is monitored very accurately using a laser interferometer. Dropping chamber of 33cm. Accuracy of approx. 2μGal.

Planes

HALO – the German High Altitude and LOng Range Research Aircraft (HALO).

GEOHALO –
Multidisciplinary geoscientific airborne mission on the HALO aircraft over Italy June 2012

GFZ's airborne gravimeter on HALO: CHEKAN - AM

Satellites

GRACE Note that GRACE consists of two satellites which are in a low orbit and the distance between them is accurately measured. Changes in this separation are caused by increases and decreases in gravity.
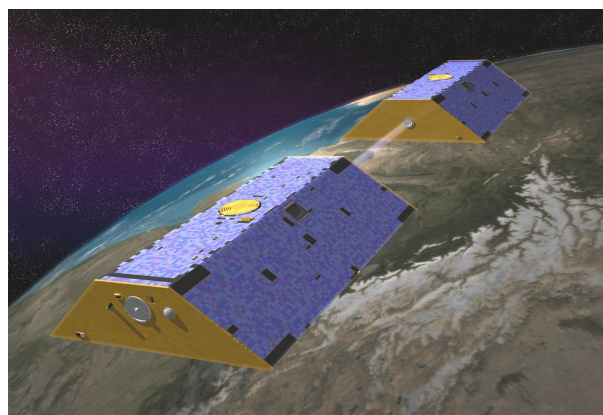
https://upload.wikimedia.org/wikipedia/commons/e/e6/GRACE_artist_concept.jpg

Examples of applications using GRACE data:

---

[17]http://microglacoste.com/product/fg5-x-absolute-gravimeter/

- Inference of mantle viscosity from GRACE and relative sea level data [1043]

- Exploring the uncertainty in GRACE estimates of the mass redistributions at the Earth surface: implications for the global water and sea level budgets [103]

GOCE   Gravity Field and Steady-State Ocean Circulation Explorer (GOCE) was the first of ESA's Living Planet Programme satellites intended to map in unprecedented detail the Earth's gravity field with a spatial resolution up to 80 km. The spacecraft's primary instrumentation was a highly sensitive gravity gradiometer consisting of three pairs of accelerometers which measured gravitational gradients along three orthogonal axes.
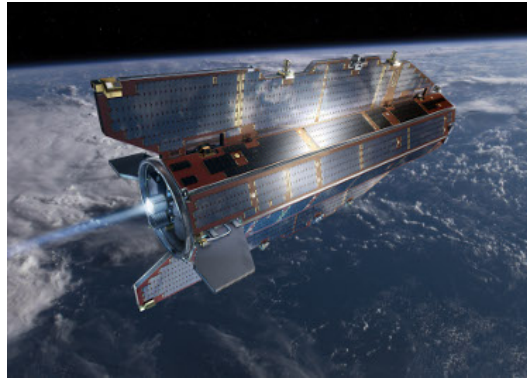
Examples of applications using GOCE data:

- GOCE gravitational gradients along the orbit [129]

- Moho Estimation Using GOCE Data: A Numerical Simulation [1124]

- Global Moho from the combination of the CRUST2.0 model and GOCE data [1125]

- Advancements in satellite gravity gradient data for crustal studies [382]

- Sensitivity of GOCE Gravity Gradients to Crustal Thickness and Density Variations: Case Study for the Northeast Atlantic Region [381]

- Mapping the mass distribution of Earth's mantle using satellite-derived gravity gradients [1030]

- GOCE gravity gradient data for lithospheric modeling [130]

- Exploration of tectonic structures with GOCE in Africa and across-continents [139]

- GEMMA: An Earth crustal model based on GOCE satellite data [1123]

- GOCE data, models, and applications: A review [910]

- Geological units and Moho depth determination in the Western Balkans exploiting GOCE data [1175]

- The combined inversion of seismological and GOCE gravity data: New insights into the current state of the Pacific lithosphere and upper mantle [1361]

17.13   Gravity anomalies

Caves and cavities

Underground man-made structures (bunkers)

Mineral deposits

Impact craters

Salt layers

17.14   Gravity reductions

In gravity work, more than in any other branch of geophysics, large and (in principle) calculable effects are produced by sources which are not of direct geological interest. These effects are removed by reductions involving sequential calculation of a number of recognized quantities.

Latitude correction   Because of the shape of the Earth, latitude has a large effect on gravity measurements, as visible in the following figure:

gravity as measured by V. Meinesz onboard the K-XVIII submarine. On the left: chronological order.

On the right: same measurements, but organised per latitude.
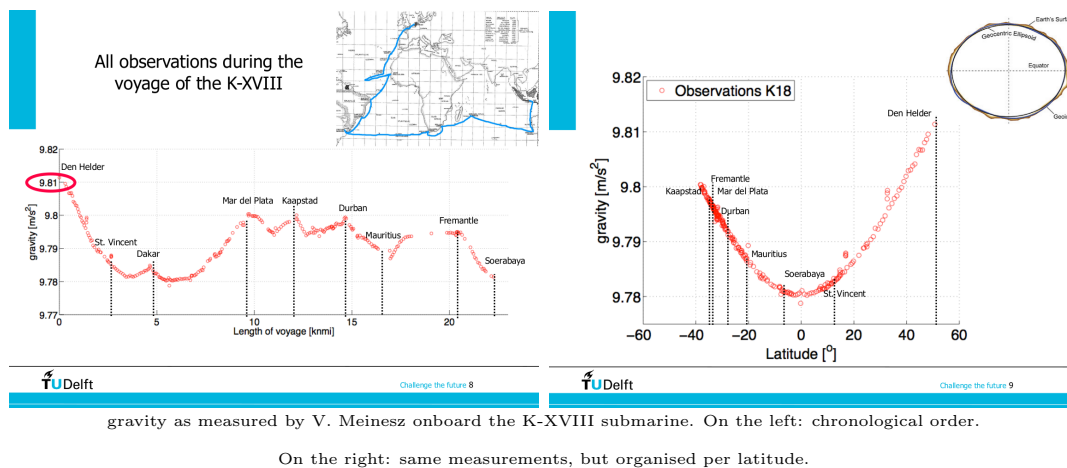
It is obvious that we are not interested in this long wavelength pattern, but rather in the deviations from it.

The formula is as follows:

$$g_n = 978031.85(1.0 + 0.005278895\sin^2(lat) + 0.000023462\sin^4(lat))(\text{mgal})$$

where lat is the latitude.

Free-air correction   After substracting the above signal, the observed gravity will be due in part to the height of the gravity station above the sea-level reference surface. An increase in height implies an increase in distance from the Earth's centre of mass and the effect is negative for stations above sea level ($g \propto r^{-2}$). The free-air correction is thus positive and the quantity obtained after applying both the latitude and free-air corrections is termed the free-air anomaly or free-air gravity.

Bouguer correction

Terrain correction   material10.pdf

17.15 How not to think about gravity (or Earth Sciences)

I think I have a crush on Beyoncé

Whatever floats your boat

No, thats Buoyancy

jeff · 1 year ago

As for gravity one simple question to debunk gravity because the example you gave is density. The heavier something is the harder it falls thats density not gravity. So now to debunk gravity dumbass.
How can gravity hold 350 trillion gallons of water to a spinning ball yet gravity can let a fly roam free or when my dog shakes its wet fur why does the water fly everywhere and not stick to the ground ??????
Thats common sense not your dumbass shit.

YOU DON'T BELIEVE IN GRAVITY?
GO JUMP OFF A CLIFF AND SEE WHAT HAPPENS

OKAY.. THAT ONLY PROVED I AM DENSER THAN THE AIR SO I FALL THROUGH IT, BUT NOT AS DENSE AS THE WATER SO I FLOAT IN IT
SCIENCE CLASS DISMISSED!
GRAVITY IS JUST DENSITY

Sai
So gravity is really a lie ?
on Sat   Like   Reply   More

Tyler
Not a lie but disorder.... disfunction.... the atoms are shooting all over the place and poles point in any direction.
on Sat   Like   Reply   More

Thomas
Gravity exist but not as we know it. Gravity is an electrical phenomenon called incoherent dielectric acceleration in the form of universal compression radiating from the black sun 😊

HE DIED OF NATURAL CAUSES.

YOU PUSHED HIM OFF A ROOF.

GRAVITY IS NATURAL.

Clarence
        Plate tectonics is false. Like with the magnetic pole flip, it lacks a plausible mechanism to produce the energy needed to move kajiullions of cubic miles of crust a bazillion miles, and that at a speed high enough to thrust up the Rockies and the Himalayas. I don't buy it.

I don't believe in magic.
9 hrs   Like   Reply   More

EARTH is a realm, it is not a planet. It is not an object, therefore, it has no edge. Earth would be more easily defined as a system environment. Earth is also a machine, it is a Tesla coil. The sun and moon are powered wirelessly with the electromagnetic field (the Aether). This field also suspends the celestial spheres with electo-magnetic levitation. Electromagnetic levitation disproves gravity because the only force you need to counter is the electromagnetic force, not gravity. The stars are attached to the FIRMAMENT.

Kenny
7 hrs · Facebook for Android

I always wondered if aliens are living on the other side of the disc. Maybe they don't know they are upside down. 🤣🤣🤣

F Gravity can be debunked with one statement. No matter in this universe can create uniform force because matter is always irregular (even in earth core if it exists it would be fluid form because of the pressure applied). When feather and steel ball fall at same acceleration in vaccum shows the gravity effect on the object is uniform which is impossible with any matter. My only conclusion is all matter inthis universe is falling at same speed if earth is flat. All my senses tell me earth is infinite horizon. Even if it is a potato then gravity can only be buoyancy effect of matter occupy the space and all the objects try to get to the center of that space. That means gravity varies bases on size of the planet not on the mass of the planet.

Free Speech • 2 days ago

P Gravity is 'their' top trump. I personally believe there is a little more to it than density alone, something to do with the speed of light, a reciprocal perhaps. I also think we are upside down, in the anti matter cycle if you prefer. The wrong side of eternity.

Paul • 3 days ago (edited)

👍   👎   💬   ⋮

Jacqui 20 hrs

Has there been any experiment that proves that water can stick to a sphere? If so, I'd like to see it. Also, quick question: if "gravity" can cause the oceans to adhere to the globe, wouldn't the same laws apply to a small ball in front of me? I've tried sticking water to one but it just ain't happening? Do the laws of gravity only apply to large balls? And if so, at what point are balls considered large? (No vulgar replies please - haha!)

👍 Like   💬 Comment

😮 10

Cindy To my knowledge there is no repeatable experiment to prove water sticks to a spinning ball of any size.
Like · Reply · 2 · 19 hrs

Jerry Maybe in a vacuum chamber? I imagine, in a vacuum, the water would pool up around the ball and the ball would just spin freely underneath it while the water stays in the same position.
Like · Reply · 3 · 19 hrs · Edited

Jacqui Is there a video of this experiment?
Like · Reply · 1 · 19 hrs

Jerry I don't know, I just made that hypothesis. I've thought about it before, but I don't think there's been an experiment performed.
Like · Reply · 19 hrs · Edited

Jacqui Jerry - do one and video it!
Like · Reply · 1 · 19 hrs

Jerry It's a good idea, but it think it's beyond my capability of time and resources. I have a son coming to this earth next month, so he's going to be my priority for a while. Maybe one day.
Like · Reply · 3 · 19 hrs

Cindy I would say let's demand an experiment from the scientific community that swears by gravity.
Like · Reply · 4 · 19 hrs

Cindy They won't because they can't.
Like · Reply · 3 · 18 hrs

John its all down to relative mass isn't it? and space is supposedly a vacuum, if it exists.
Like · Reply · 3 · 14 hrs

Tippy Gravity doesn't exsist,water forms to it's container,unless ball is frozen, lol 😝
Like · Reply · 1 · 14 hrs

We all agree Gravity is nonsense but... Here's my question: is Density and Buoyancy enough? Or is magnetism at play as well?

Reece Density and buoyancy are the effect, electromagnetism is the cause. All in line with the Torus field.
2 hrs Like Reply More   👍3

Tanakah You said correctly... Magnetic fields all around us.. Gravity is a scam
3 hrs Like Reply More   👍2

Lew I don't think buoyancy and density is enough since it doesn't explain the reason things "fall" "down". Incoherent magnetism, dielectric acceleration and electrostatics do explain that in my opinion.
3 hrs Like Reply More   👍5

Mark EMF plays a vital role on our Plane Flat Earth's Realm, movement of luminaries, light (visible and non visible), energy, eclipses, tides, spectrums, aether, the light described in the Bible prior on creation of luminary Sun, plane of inertia, line and point of inertia of this Realm, torus field and a lot more. All are true science that God created. 🙏
3 hrs Like Reply More   👍😮5

Mark I agree with your thoughts on this. I feel that density and buoyancy are enough, however very possible that magnetism comes into play somehow. But gravity is definitely a bunch of b.s., they need gravity to explain their claims of objects orbiting other objects. It seems to be at once both a force of adhesion and a force of suspension and rotation. It's the Swiss knife go to reason used for any claims that they cannot explain.
3 hrs Like Reply More   👍1

Jessie Aether... 👍2
3 hrs Like Reply More

Eddie Even the smallest of creatures, like the bumblebee, use magnetism to hover/fly. Resonance of frequency within the aether, achieves this.
3 hrs Like Reply More   👍1

Bert Yes magnetism steers the tide
3 hrs Like Reply More

Ray Water is not magnetic
45 mins Like Reply More

Mike Gravity is a word to misguide us so the true power of our world is never discovered. Would it eventually destroy us once known on a massive scale, Possibly, as those who occupied our flat earth before us are gone with no history that we know of or are told.
2 hrs Like Reply More   👍1

NO MATTER HOW MUCH YOU ADD SUBTRACT, MULTIPLE OR DIVIDE

WATER WILL NEVER STICK TO A BALL
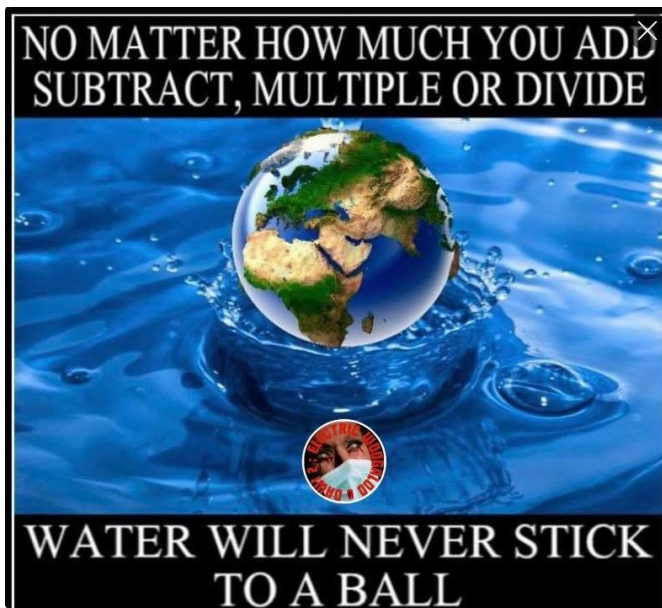
Americans when Newton discovered gravity :

Sam
Gravity is a cover-up concept for buoyancy and density.

last Sun   Like   More

Keanu Reeves and Bernie Sanders are sitting 1.7 m apart. Keanu has a mass of 76 kg. Bernie has a mass of 64 kg.

(a) What is the gravitational force that Keanu exerts on Bernie?

[ ] N

(b) What is the gravitational force that Bernie exerts on Keanu?

[ ] N
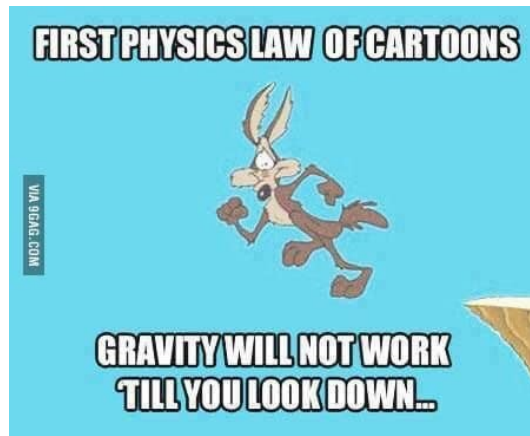
17.16 Solutions to exercises of GEO3-1313

17.16.1 Problem 1
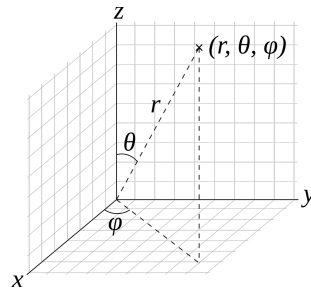
$$
\begin{aligned}
I &= \frac{1}{3}(I_x + I_y + I_z) \\
&= \frac{1}{3}\iiint_V \rho(r)(y^2 + z^2)dV + \frac{1}{3}\iiint_V \rho(r)(x^2 + z^2)dV + \frac{1}{3}\iiint_V \rho(r)(x^2 + y^2)dV \\
&= \frac{1}{3}\iiint_V \rho(r)(2x^2 + 2y^2 + 2z^2)dV \\
&= \frac{2}{3}\iiint_V \rho(r)r^2 dV \\
&= \frac{2}{3}\iiint_V \rho(r)r^2 r^2 \sin\theta dr\, d\theta\, d\phi \\
&= \frac{2}{3}\int_0^R \int_0^\pi \int_0^{2\pi} \rho(r)r^2 r^2 \sin\theta dr\, d\theta\, d\phi \\
&= \frac{2}{3}\left(\int_0^\pi \sin\theta d\theta\right)\left(\int_0^{2\pi} d\phi\right)\int_0^R \rho(r)r^4 dr \\
&= \frac{8\pi}{3}\int_0^R \rho(r)r^4 dr \tag{17.132}
\end{aligned}
$$

Alternative solution: I is evaluated for the special case where the rotation axis is the z-axis, where $d = r\sin\theta$. Substitution in $I = \int_V \rho d^2 dV$ yields

$$
\begin{aligned}
I &= \iiint_V \rho(r)(r^2 \sin^2\theta)dV \\
&= \int_0^R \int_0^\pi \int_0^{2\pi} \rho(r)(r^2 \sin^2\theta)r^2 \sin\theta dr\, d\theta\, d\phi \\
&= \left(\int_0^\pi \sin^3\theta d\theta\right)\left(\int_0^{2\pi} d\phi\right)\int_0^R \rho(r)r^4 dr
\end{aligned}
$$

By writing $\sin^3\theta = \sin\theta(1 - \cos^2\theta)$ we can operate a change of variables $u = \cos\theta$, and we arrive at $\int \sin^3\theta d\theta = 4/3$ (skipping a few steps here) and we obtain the expected result.

17.16.2   Problem 2

For a uniform sphere, we have $\rho(r) = \rho_0$. Then

$$
\begin{aligned}
I &= \frac{8\pi}{3} \int_0^R \rho_0 r^4 dr \\
&= \frac{8\pi}{3} \rho_0 \int_0^R r^4 dr \\
&= \frac{8\pi}{15} \rho_0 R^5 \int_0^R r^4 dr
\end{aligned}
$$

The mass of the sphere is

$$
\begin{aligned}
M &= \iiint_V \rho(r) dV \\
&= \rho_0 \int_0^R \int_0^\pi \int_0^{2\pi} r^2 \sin\theta dr\, d\theta\, d\phi \\
&= \rho_0 \frac{4}{3}\pi R^3 \tag{17.133}
\end{aligned}
$$

In the end we get

$$
I = \frac{2}{5}MR^2
$$

When all the mass is concentrated in the center, then $\rho(r) = \rho_0 \delta(r)$ where δ is the Dirac delta function[18]. Then

$$
\begin{aligned}
I &= \frac{8\pi}{3} \int_0^R \rho_0 \delta(r) r^4 dr \\
&= \frac{8\pi}{3} \rho_0 \int_0^R \delta(r) r^4 dr \\
&= 0 \tag{17.134}
\end{aligned}
$$

When all the mass is concentrated in a shell of zero thickness of radius R, then $\rho(r) = \rho_0 \delta(r - R)$, so

$$
\begin{aligned}
I &= \frac{8\pi}{3} \int_0^R \rho_0 \delta(r - R) r^4 dr \\
&= \frac{8\pi}{3} \rho_0 \int_0^R \delta(r - R) r^4 dr \\
&= \frac{8\pi}{3} \rho_0 R^4 \tag{17.135}
\end{aligned}
$$

Conversely, its mass is

$$
\begin{aligned}
M &= \iiint_V \rho(r) dV \\
&= \rho_0 \int_0^R \int_0^\pi \int_0^{2\pi} \delta(r - R) r^2 \sin\theta dr\, d\theta\, d\phi \\
&= \rho_0 4\pi R^2 \tag{17.136}
\end{aligned}
$$

and then

$$
I = \frac{2}{3}MR^2
$$

---

[18]https://en.wikipedia.org/wiki/Dirac_delta_function

17.16.3 Problem 3

$$\langle \rho \rangle = \frac{1}{V} \iiint_V \rho dV$$

$$= \frac{1}{\frac{4}{3}\pi R^3} \iiint_V \rho(r) r^2 \sin\theta dr \, d\theta d\phi$$

$$= \frac{1}{\frac{4}{3}\pi R^3} 4\pi \int_0^R \rho(r) r^2 dr$$

$$= \frac{3}{R^3} \int_0^R \rho(r) r^2 dr \tag{17.137}$$

We then turn to the moment of inertia:

$$I = \frac{8\pi}{3} \int_0^R \rho(r) r^4 dr \tag{17.138}$$

$$= f M R^2 \tag{17.139}$$

where

$$M = \iiint_V \rho(r) dV = V \underbrace{\frac{1}{V} \iiint_V \rho(r) dV}_{\langle \rho \rangle} = \frac{4}{3}\pi R^3 \langle \rho \rangle$$

We then insert this expression of M in Eq. (17.139):

$$I = f \frac{4}{3}\pi R^3 \langle \rho \rangle R^2 = f \frac{4}{3}\pi R^5 \langle \rho \rangle$$

Equating this to Eq. 17.138 yields

$$\frac{8\pi}{3} \int_0^R \rho(r) r^4 dr = f \frac{4}{3}\pi R^5 \langle \rho \rangle$$

or,

$$f \langle \rho \rangle R^5 = 2 \int_0^R \rho(r) r^4 dr \tag{17.140}$$

We now make use of the expression for the density:

$$\rho(r) = \begin{cases} \rho_c & 0 \leq r \leq R_c \\ \rho_m & R_c \leq r \leq R \end{cases}$$

Then Eq. (17.137) yields

$$\langle \rho \rangle = \frac{3}{R^3} \int_0^R \rho(r) r^2 dr$$

$$= \frac{3}{R^3} \left(\int_0^{R_c} \rho_c r^2 dr + \int_{R_c}^R \rho_m r^2 dr \right)$$

$$= \frac{3}{R^3} \left(\frac{R_c^3}{3} \rho_c + \frac{1}{3}(R^3 - R_c^3)\rho_m \right)$$

$$= \frac{R_c^3}{R^3} \rho_c + (1 - \frac{R_c^3}{R^3})\rho_m$$

$$= \left(\frac{R_c}{R}\right)^3 \rho_c + \left[1 - \left(\frac{R_c}{R}\right)^3\right] \rho_m \tag{17.141}$$

We know ρ_m, but not ρ_c, so we write:

$$\rho_c = \rho_m \left[1 + \left(\frac{R}{R_c} \right)^3 \left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right) \right] \tag{17.142}$$

We now turn to Eq. (17.140). Since

$$
\begin{aligned}
2 \int_0^R \rho(r) r^4 dr &= 2 \int_0^{R_c} \rho_c r^4 dr + 2 \int_{R_c}^R \rho_m r^4 dr \\
&= \frac{2}{5} \left[R_c^5 \rho_c + (R^5 - R_c^5) \rho_m \right]
\end{aligned}
$$

then

$$f \langle \rho \rangle R^5 = \frac{2}{5} \left[R_c^5 \rho_c + (R^5 - R_c^5) \rho_m \right]$$

or,

$$\frac{5 f \langle \rho \rangle R^5}{2 \rho_m} = R_c^5 \frac{\rho_c}{\rho_m} + (R^5 - R_c^5)$$

$$\frac{5 f \langle \rho \rangle R^5}{2 \rho_m} = R_c^5 (\frac{\rho_c}{\rho_m} - 1) + R^5)$$

Now dividing by R^5 on each side:

$$\frac{5 f \langle \rho \rangle}{2 \rho_m} = \left(\frac{R_c}{R} \right)^5 \left(\frac{\rho_c}{\rho_m} - 1 \right) + 1$$

Using Eq. (17.142), we can write

$$\frac{\rho_c}{\rho_m} - 1 = \left(\frac{R}{R_c} \right)^3 \left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right)$$

so finally

$$\frac{5 f \langle \rho \rangle}{2 \rho_m} - 1 = \left(\frac{R_c}{R} \right)^5 \left(\frac{R}{R_c} \right)^3 \left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right)$$

or,

$$\frac{5 f \langle \rho \rangle}{2 \rho_m} - 1 = \left(\frac{R_c}{R} \right)^2 \left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right)$$

and then

$$\left(\frac{R_c}{R} \right) = \left(\frac{\frac{5 f \langle \rho \rangle}{2 \rho_m} - 1}{\left(\frac{\langle \rho \rangle}{\rho_m} - 1 \right)} \right)^{1/2}$$

17.16.4 Problem 4

Assume a uniform mantle ρ_m and core ρ_c. For the total mass we have

$$M = \int_V \rho dV = 4\pi \int_0^R \rho(r)r^2 dr = \frac{4\pi}{3}R_c^3\rho_c + \frac{4\pi}{3}(R^3 - R_c^3)\rho_m$$

For the moment of inertia we have,

$$I = \frac{8\pi}{3}\int_0^R \rho(r)r^4 dr = \frac{8\pi}{15}R_c^5\rho_c + \frac{8\pi}{15}(R^5 - R_c^5)\rho_m$$

In matrix form the above equations become:

$$\left(\begin{array}{cc} \frac{4\pi}{3}R_c^3 & \frac{4\pi}{3}(R^3 - R_c^3) \\ \frac{8\pi}{15}R_c^5 & \frac{8\pi}{15}(R^5 - R_c^5) \end{array} \right) \cdot \left(\begin{array}{c} \rho_c \\ \rho_m \end{array} \right) = \left(\begin{array}{cc} M & I \end{array} \right)$$

We use Cramers rule

$$\left(\begin{array}{cc} a11 & a12 \\ a21 & a22 \end{array} \right) \cdot \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left(\begin{array}{c} b_1 \\ b_2 \end{array} \right) \qquad \Rightarrow \qquad \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \frac{1}{\Delta}\left(\begin{array}{cc} a22 & -a12 \\ -a21 & a11 \end{array} \right) \cdot \left(\begin{array}{c} b_1 \\ b_2 \end{array} \right)$$

In our case the determinant of the matrix is

$$\Delta = \frac{32\pi^2}{45}\left[R_c^3(R^5 - R_c^5) - R_c^5(R^3 - R_c^3) \right]$$

17.16.5 Problem 5

skipped

17.16.6 Problem 6

see lecture notes

17.16.7 Problem 7

$$g = \frac{GM}{R^2} = \frac{6.67e-11 * 5.97e24}{6371000^2} \simeq 9.8$$

17.16.8 Problem 8

$$p_{CMB} = \int \rho g dz \simeq \rho_0 g_0 (R_{Earth} - R_{CMB}) \simeq 127.5 GPa$$

17.16.9   Problem 9

$$U(\vec{r}) = -\frac{\mathcal{G}m_1}{|\vec{r}_1 - \vec{r}|} = -\frac{\mathcal{G}m_1}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}}$$

$$\vec{\nabla}U = \begin{pmatrix} \partial_x U \\ \partial_y U \\ \partial_z U \end{pmatrix}$$

We have

$$\begin{aligned}
\partial_x U &= -\mathcal{G}m_1 \frac{\partial}{\partial x} \frac{1}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}} \\
&= -\mathcal{G}m_1 \cdot -\frac{1}{2} \frac{-2(x_1 - x)}{[(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2]^{3/2}} \\
&= \mathcal{G}m_1 \frac{1}{[(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2]} \frac{-(x_1 - x)}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}} \\
&= \mathcal{G}m_1 \frac{1}{|\vec{r}_1 - \vec{r}|^2} \frac{-(x_1 - x)}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}}
\end{aligned}$$

We repeat this operation for ∂_y and ∂_z and finally:

$$-\vec{\nabla}U = \begin{pmatrix} \partial_x U \\ \partial_y U \\ \partial_z U \end{pmatrix} = \mathcal{G}m_1 \frac{1}{|\vec{r}_1 - \vec{r}|^2} \begin{pmatrix} \frac{(x_1 - x)}{|\vec{r}_1 - \vec{r}|} \\ \frac{(y_1 - y)}{|\vec{r}_1 - \vec{r}|} \\ \frac{(z_1 - z)}{|\vec{r}_1 - \vec{r}|} \end{pmatrix} = \frac{\mathcal{G}m_1}{|\vec{r}_1 - \vec{r}|^2} \vec{e}_{\vec{r}_1 \vec{r}} = \vec{g}$$

17.16.10   Problem 10

$$\int_V \Delta U\, dV = \int_V 4\pi \mathcal{G}\rho\, dV = 4\pi \mathcal{G} \int_V M\delta(\vec{r} - \vec{r}_0)\, dV = 4\pi \mathcal{G}M$$

$$\int_V \Delta U\, dV = \int_V \vec{\nabla}^2 U\, dV = \int_V \vec{\nabla} \cdot \vec{\nabla}U\, dV = \int_\Gamma \vec{\nabla}U \cdot \vec{n}\, dS = \int_\Gamma (-\vec{g}) \cdot \vec{n}\, dS = \int_\Gamma g\, dS = 4\pi r^2 g$$

Note that we have $\vec{g}$ which is pointing towards the center and therefore has the opposite direction to $\vec{n}$ which is normal to the shell surface so that $(-\vec{g}) \cdot \vec{n} = g$. Also the integral on Γ is at constant radius so $g(r)$ can be taken out of the integral.

Finally we obtain

$$g = \frac{\mathcal{G}M}{r^2}$$

17.16.11   Problem 11

We start from $\vec{\nabla}^2 U = 4\pi \mathcal{G}\rho$. We then have

$$[\vec{\nabla}^2][U] = [\mathcal{G}][\rho]$$

so

$$[U] = [\mathcal{G}][\rho]/[\vec{\nabla}^2] = M^{-1}L^3 T^{-2} \cdot ML^{-3} \cdot L^2 = L^2 T^{-2} = \underbrace{ML^2 T^{-2}}_{energy}/M$$

(see lecture notes on physical dimensions)

17.16.12   Problem 12

Escape velocity is speed at which kinetic energy is equal to gravitational potential energy, i.e.

$$\frac{1}{2}mv^2 = mgH$$

so $v = \sqrt{2gH}$ and since $g = \mathcal{G}M/R^2$ then the escape velocity at the surface (i.e. $H = R$) is given by

$$v = \sqrt{2\mathcal{G}M/R}$$

$$v_{earth} \simeq 11.2km/s$$
$$v_{moon} \simeq 2.4km/s$$

See https://www.newworldencyclopedia.org/entry/Escape_velocity

17.16.13   Problem 13

$$
\begin{aligned}
E &= -\int_V \rho U dV \\
&= -\int_V \rho_0 U(r) r^2 \sin\theta \, dr d\theta d\phi \\
&= -4\pi\rho_0 \int_0^R r^2 U(r) dr \\
&= -4\pi\rho_0 \int_0^R r^2 \left[\frac{2\pi}{3}\mathcal{G}\rho_0 r^2 - \frac{3}{2}\frac{\mathcal{G}M}{R} \right] dr \\
&= -\frac{8\pi^2 \rho_0^2}{3}\mathcal{G} \int_0^R r^4 dr + 6\rho_0\pi\frac{\mathcal{G}M}{R} \int_0^R r^2 dr \\
&= \dots \\
&= \frac{8\pi}{5}\rho_0 M R^2 \mathcal{G}
\end{aligned}
\tag{17.143}
$$

17.16.14   Problem 14

We start from the Laplace operator in spherical coordinates:

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}$$

Because of the symmetry of the problem, the solution is expected to only depend on r, and not on θ nor ϕ, so that $\partial_\theta \to 0$ and $\partial_\phi \to 0$ in the equation above. We end up with:

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)$$

Inside the planet, the density is not zero, so we need to solve a Poisson equation

$$\Delta U = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 4\pi\mathcal{G}\rho_0$$

Outside the planet the density is zero and we need to solve a Laplace equation:

$$\Delta U = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 0$$

We start with the Poisson equation which we rewrite as follows:

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = 4\pi\mathcal{G}\rho_0 r^2$$

We integrate once and obtain

$$r^2\frac{\partial U}{\partial r} = \frac{4\pi}{3}\mathcal{G}\rho_0 r^3 + A$$

where A is a constant to be specified later. We divide by r^2 and obtain

$$\frac{\partial U}{\partial r} = \frac{4\pi}{3}\mathcal{G}\rho_0 r + \frac{A}{r^2}$$

The radial component of the gradient operator is simply ∂_r so that the equation above is (save a minus sign) g_r:

$$g_r(r) = -\frac{4\pi}{3}\mathcal{G}\rho_0 r - \frac{A}{r^2}$$

When $r \to 0$ the gravity acceleration must remain finite so we need to set $A = 0$. Then

$$\boxed{g_r(r)|_{inside} = -\frac{4\pi}{3}\mathcal{G}\rho_0 r}$$

We integrate once more and obtain

$$U(r)|_{inside} = \frac{2\pi}{3}\mathcal{G}\rho_0 r^2 + B$$

where B is a constant.

We now turn to the Laplace equation outside the planet which yields

$$r^2\frac{\partial U}{\partial r} = C$$

where C is a constant. Then it follows that

$$\frac{\partial U}{\partial r} = \frac{C}{r^2}$$

or,

$$U(r) = -\frac{C}{r} + D$$

When $r \to \infty$ the potential tends to zero, so that $D = 0$. Then

$$\boxed{U(r)_{outside} = -\frac{C}{r}}$$

and from $g_r(r) = -\partial_r U$ we get

$$g_r(r)|_{outside} = -\frac{C}{r^2}$$

We have solved the Poisson and Laplace equations but remain two constants to be specified. In order to do so we invoke the continuity of the gravity acceleration and potential at the surface of the planet:

$$g_r(r = R)|_{inside} = g_r(r = R)|_{outside}$$
$$U(r = R)|_{inside} = U(r = R)|_{outside}$$

The first continuity condition yields

$$-\frac{C}{R^2} = -\frac{4\pi}{3}\mathcal{G}\rho_0 R$$

i.e., $C = M\mathcal{G}$. The second continuity condition then yields

$$-\frac{C}{R} = -\frac{M\mathcal{G}}{R} = \frac{2\pi}{3}\mathcal{G}\rho_0 R^2 + B$$

i.e. $B = -\frac{3}{2}\frac{M\mathcal{G}}{R}$.

If all the mass is concentrated at the origin then by definition

$$g_r(r) = \frac{M\mathcal{G}}{r^2} \qquad U(r) = -\frac{M\mathcal{G}}{R}$$

Finally

$$P(r) = -\int_r^R \rho(r')g(r')dr' = \int_r^R \rho_0\frac{4\pi}{3}\mathcal{G}\rho_0 r'dr' = \frac{2\pi}{3}\mathcal{G}\rho_0^2(R^2 - r^2)$$

17.16.15 Problem 15

We start from (17.36), i.e

$$U(r) = -\int_r^\infty \frac{Gm(r')}{r'^2}dr'$$

Outside the sphere, $r > R$ and the mass at any location $r' > R$ is simply M. Then

$$U(r) = -\int_r^\infty \frac{GM}{r'^2}dr' = -\frac{\mathcal{G}M}{r}$$

When $r < R$ we can split the integral in two:

$$U(r) = -\int_r^R \frac{Gm(r')}{r'^2}dr' - \int_R^\infty \frac{Gm(r')}{r'^2}dr'$$

We have just computed the second term so we focus on the first one. In this integral the mass $m(r') = \frac{4}{3}\pi r'^3\rho_0$ so that

$$U(r) = -\int_r^R \frac{G}{r'^2}\frac{4}{3}\pi r'^3\rho_0 dr' - \frac{\mathcal{G}M}{r} = -\frac{1}{2}\frac{4}{3}\pi\mathcal{G}(R^2 - r^2) - \frac{\mathcal{G}M}{r} = \frac{2\pi}{3}\rho_0\mathcal{G}r^2 - \frac{3}{2}\frac{\mathcal{G}M}{r}$$

Chapter 18

Mantle Dynamics GEO4-1416 syllabus (Utrecht University)

Remark. *This handout was written by W. Spakman and is for a large part based on a syllabus by Dr. A.P. van den Berg and Prof. N.J. Vlaar and on material from the book "Mantle convection in the Earth and Planets" by Schubert, Turcotte, and Olson, Cambridge University Press, 2002.*

Additional resources:
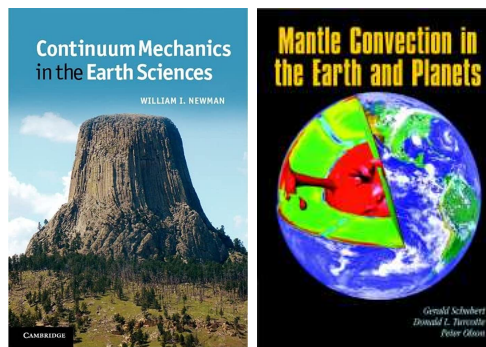
Left: William I Newman. *Continuum Mechanics in the Earth Sciences.* Cambridge University Press, 2012. ISBN: 978-0-521-56289-8; Right: G. Schubert, D.L. Turcotte, and P. Olson. *Mantle Convection in the Earth and Planets.* Cambridge University Press, 2001. ISBN: 0-521-70000-0. DOI: 10.1017/CBO9780511612879.

In what follows (and in the entire fieldstone document) vectors are denoted by an arrow, e.g. $\vec{v}$, while tensors are denoted by a bold font, e.g. $\boldsymbol{\sigma}$. It is also my experience that each book/syllabus uses its own set of notations. In particular, and rather to my own confusion, some books call the full stress tensor $\boldsymbol{\tau}$ and not $\boldsymbol{\sigma}$, and call the deviatoric stress $\boldsymbol{\sigma}$. Some sources make no difference between the viscous stress tensor and the deviatoric stress tensor. Also it happens (especially in the mathematical literature) that the strain rate tensor is called $\boldsymbol{D}$ instead of $\boldsymbol{\varepsilon}$ or $\dot{\boldsymbol{\varepsilon}}$, and that the stress tensor is called $\boldsymbol{T}$. All in all be careful when reading additional sources.

Continuum mechanics	↔	continuummechanica
Coordinate system	↔	coordinatenstelsel
Cartesian coordinates system	↔	cartesische coordinatenstelsel
Scalar	↔	scalair, scalaire grootheid of scalar
Vector	↔	vector
Tensor	↔	tensor
Axis	↔	axis
Unit vector	↔	eenheidsvector
Cross product	↔	vectorprodukt of uitprodukt
volume forces	↔	volumekrachten
Surface forces	↔	oppervlaktekrachten
Stress tensor	↔	spanningstensor

18.0.1 The continuity equation

18.1 Review of some essentials of continuum mechanics

Newtonian mechanics deal with particles and rigid (undeformable) bodies on which forces are acting. The application of Newtonian mechanics to realistic media (gases, fluids, solids) is undoable simply because of the many particles (atoms, molecules) involved. Continuum mechanics tackles this problem by assuming that physical fields (e.g. density, temperature, velocity) can be viewed as (piece-wise) continuous functions defined on the time and space coordinates involved in the description of macroscopic matter. The idea is essentially that a tiny cube with sides of, say, 10^{-8} m already contains a sufficient number of atoms (millions) which allows for establishing physically meaningful descriptions of quantities as temperature and density of the cube. In continuum mechanics we are mostly interested in material behavior on much larger scales than 10^{-8} m for which is assumed that physical quantities are smooth functions of time and spatial coordinates.

The forces involved in the deformation of a continuum are postulated to be **body forces** $\vec{b}$ [N m^{-3}], such that $\vec{b}dV$ is the force acting on the infinitesimal volume dV, and surface **tractions** $\vec{t}^{\vec{n}}$ [N m^{-2}] (e.g. internal friction, applied surface tractions), such that $\vec{t}^{\vec{n}}dS$ is the force acting on the infinitesimal surface dS. It is usual to write $\vec{b} = \rho\vec{g}$, with ρ [kg m^{-3}] the mass density and with $\vec{g}$ [m s^{-2}] the acceleration due to the body force, which in mantle dynamics is gravity.

$\vec{b}$ and $\vec{t}^{\vec{n}}$ are force densities which after integration over a volume or a surface, respectively, lead to net forces acting on the volume or surface. The traction (or stress vector) is defined as

$$\vec{t}^{\vec{n}} = \lim_{\Delta S \to 0} \frac{\sum_i f_i^{\vec{n}}}{\Delta S},$$

which expresses the force per unit area working on a tiny surface ΔS with unit normal $\vec{n}$ (defining the orientation of the surface). The forces $f_i^{\vec{n}}$ can be viewed as the atomic forces [N] that are applied at the $\vec{n}$-side of ΔS to atoms at the other side of the surface. To maintain equilibrium, by the third law of Newton, the traction applied to the $-\vec{n} - side$ of the surface is $\vec{t}^{-\vec{n}} = -\vec{t}^{\vec{n}}$. Tractions depend on the orientation of the surface. In principle, one can draw an infinite number of oriented surfaces through one point, each associated with a different traction.

Tractions are usually separated into the thermodynamic pressure force $p\vec{n}$ and the traction $\vec{\tau}$ related to mechanical deformation: $\vec{t}^{\vec{n}} = -p\vec{n} + \vec{\tau}$ $(p > 0)$. The thermodynamic pressure (a traction always acting perpendicular to any surface) is obtained from the equation of state $f(\rho, p, T)$ relating thermodynamic quantities density, pressure, and temperature of a continuum. The sign convention in continuum mechanics is that **compression is negative and tension is positive** (in geology this is usually the other way around).

18.1.1 Stress

Stress is a second order tensor quantity, which is defined by the following steps:

1. Assume a Cartesian coordinate frame in a point of interest for which the coordinate axis are spanned by three unit orthogonal vectors $\vec{e}_i$ $(i = 1, 2, 3)$,

2. Imagine a tiny cube centered about the origin and with its faces parallel to the coordinate planes,

3. Consider the 3 tractions $\vec{t}^{\vec{e}_i}$ that are acting on the three positive faces of the cube (i.e. the faces which have the normal vectors $\vec{e}_i$),

4. Lastly, define the components σ_{ij} of the stress tensor $\boldsymbol{\sigma}$ as

$$\sigma_{ij} = \vec{t}^{\vec{e}_i} \cdot \vec{e}_j \tag{18.1}$$

When the stress tensor is visualized as a matrix, the three rows are the tractions on the positive faces of the cube. The diagonal elements of the stress tensor $\boldsymbol{\sigma}$ are called normal stresses and the off-diagonal elements are the shear stresses.

Stress is a physical quantity, independent of coordinate frame, but the actual values of components σ_{ij} of the stress tensor can only be computed in a coordinate system. These numbers are dependent on the frame adopted like the components of a flow vector (a first order tensor) are frame dependent. Second order tensors follow (by definition) the coordinate transformation rules of 3×3 matrices. From an analysis of force and force- moment balance it can be demonstrated that the stress tensor is symmetric: $\boldsymbol{\sigma} = \boldsymbol{\sigma}^T$ or $\sigma_{ij} = \sigma_{ji}$.

An eigenvalue-eigenvector analysis leads to the three principal stresses σ_k (eigenvalues) and the corresponding three corresponding principal directions $\vec{q}_k$ (eigenvectors of unit length). The latter span three mutually orthogonal (Cartesian) axes. In the principal-axes frame the stress tensor is diagonal with the three principal stresses as diagonal elements. In the principal-axes frame the tensor components are the maximal normal stresses (tractions perpendicular to the faces of a tiny cube oriented along the principal coordinate planes) compared to the normal stresses in any other coordinate system.

An important relation (the Cauchy relation) exists between the local state of stress (i.e. the stress tensor $\boldsymbol{\sigma}$) and the traction $\vec{t}^{\vec{n}}$ acting on an (arbitrarily) oriented tiny surface ΔS with unit normal $\vec{n}$:

$$\boldsymbol{\sigma} \cdot \vec{n} = \vec{t}^{\vec{n}} \tag{18.2}$$

or in components $\sigma_{ij} n_j = t_i^{\vec{n}}$ (summation convention implied). This relation states how traction can be computed from the local stress and conversely, that from known tractions on independently oriented surfaces the stress tensor can be constructed by solving Eq. (18.2). Note that it is required that $\vec{n}$ is a unit normal, i.e. $\vec{n} \cdot \vec{n} = n_j n_j = 1$.

Exercise: 1. *Determine the tractions acting on the negative faces of a tiny cube (of which the faces are aligned with the local coordinate axes) when the stress is given. How do they compare with those of the positive faces?*

18.1.2 Force balance equation of a continuum at rest

Assume a continuum (gas, liquid, solid) at rest. In this situation the net force acting on the entire continuum is $\vec{0}$ (Newton). The sum of body forces and applied surface tractions cancel in some way. This also holds for any sub-volume V. An internal stress field may still exist as a result of the applied

forces and surface tractions. The relation between the body force, tractions, and the internal stress field is derived as follows: Consider an arbitrary sub-volume V with boundary S. Internal tractions act on the boundary S (e.g. to be determined with equation (18.2) from the internal stress field at S). The following equation postulates that the total sum of body forces

$$\int_V \rho \vec{g} dV + \int_S \vec{t}^{\vec{n}} dS = \vec{0} \tag{18.3}$$

acting on V and of tractions on S leads to a zero net force acting on V.

Substituting (18.2) in the surface integral and next applying the Divergence theorem[1] one arrives at

$$\int_V \rho \vec{g} dV + \int_V \vec{\nabla} \cdot \boldsymbol{\sigma} dV = \vec{0}$$

Because V is an arbitrary volume the integrant must equal 0, which leads to the **equilibrium equation** (or momentum conservation equation):

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \rho \vec{g} = \vec{0} \qquad \text{or,} \qquad \frac{\partial \sigma_{ij}}{\partial x_j} + \rho g_i = 0 \tag{18.4}$$

This equation holds for any point in the interior of the continuum. Any traction applied at the boundary of the continuum relates to the (local) stress through equation (18.2). Equation (18.4) states that body forces are in equilibrium with the divergence of the stress tensor. Note: Gravity implies spatial variation in stress.

A similar analysis for the equilibrium of torques leads to the symmetry of the stress tensor. In this case the equilibrium equation is

$$\int_V \rho \vec{r} \times \vec{g} dV + \int_S \vec{r} \times \vec{t}^{\vec{n}} dS = \vec{0},$$

where $\vec{r}$ is the position vector $\vec{r} = (x_1, x_2, x_3)^T$. The cross products can be written in index notation using the permutation symbol ϵ_{ijk} which equals zero if at least two indices have the same value (e.g. $\epsilon_{121} = 0$), equals $+1$ if ijk is an even permutation of 123, and equals -1 if ijk is an odd permutation of 123. This leads to the following notation of the cross product of two vectors $\vec{a} \times \vec{b} = \epsilon_{ijk} \vec{e}_i a_j b_k$ and per component $(\vec{a} \times \vec{b}) = \epsilon_{ijk} a_j b_k$ leading to the torque balance for component i:

$$\int_V \rho \epsilon_{ijk} x_j g_k dV + \int_S \epsilon_{ijk} x_j t_k^{\vec{n}} dS = 0.$$

Exercise: 2. *a) Derive equation (18.4)*
b) Using a similar approach, prove the symmetry of the stress tensor from the balance of torques (assuming that no internally applied tractions exist)
c) Derive (18.4) by considering the force balance of a tiny cube

---

18.1.3 The material derivative

For a mathematical intro see Appendix T. Consider a continuum with a three-dimensional flow field $\vec{v}(x_1, x_2, x_3, t)$ dependent on the 3 spatial coordinates x_j and time t. For any differentiable scalar function T defined on these 4 parameters we can write the total differential [2]

$$dT = \frac{\partial T}{\partial t} dt + \sum_j \frac{\partial T}{\partial x_j} dx_j \tag{18.5}$$

---

[1] https://en.wikipedia.org/wiki/Divergence_theorem
[2] See any basic textbook on Calculus.

This equation can be interpreted as follows: Consider a certain point $(\vec{r}, t)$ in which the scalar function $T(\vec{r}, t)$ has continuous partial derivatives. The infinitesimal change dT, which results from going from $(\vec{r}, t)$ to $(\vec{r} + d\vec{r}, t + dt)$ in de domain of T is given by Eq. (18.5). Importantly, $d\vec{r}$ and dt can be arbitrarily chosen (including 0 values). The total differential is at the basis of the definition of the so-called directional derivative. Differentiable functions of more than 1 variable can be differentiated in arbitrary directions (in their domain) to find their rate of change in this direction with respect to a specified parameter. Particularly, we can consider the rate of change of T with the time parameter t and (spatially) in the direction of the velocity field $\vec{v}(\vec{r}, t)$. This time- derivative is easily obtained from (18.5) by coupling dt and the spatial increment $d\vec{r}$ such that $d\vec{r} = \vec{v}dt$. Substitution in (18.5) leads to the time derivative of T in the direction of the velocity vector:

$$\frac{DT}{Dt} = \frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla}T = \frac{\partial T}{\partial t} + \sum_j v_j \frac{\partial T}{\partial x_j} \qquad \text{with} \qquad \vec{v} = \frac{d\vec{r}}{dt} \qquad (18.6)$$

which is called the material derivative of T giving the rate of change of T with time in the direction of the flow $\vec{v}(\vec{r}, t)$ at a certain point $(\vec{r}, t)$. The lhs of (18.6) treats T as a function of t only in the point $(\vec{r}(t), t)$ whereas the rhs of (18.6) shows how this can be computed from the partial derivatives of $T(\vec{r}, t)$ and the local velocity $\vec{v}(\vec{r}, t)$. The partial derivative $\partial T / \partial t$ gives the temporal rate of change at fixed position $\vec{r}$, while the second term gives the spatial contribution at fixed time, i.e $\vec{v} \cdot \vec{\nabla}T$ expresses the advective contribution (carried with the flow) to DT/Dt.

Without reference to a particular scalar function the material derivative (operator) is:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla} \qquad (18.7)$$

The material derivative holds for any scalar function, particularly, for the components v_i of the velocity field leading to the particle acceleration:

$$\frac{D\vec{v}}{Dt} = \frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \vec{\nabla}\vec{v} \qquad (18.8)$$

Note that if the velocity field is time-stationary, i.e. $\partial_t = 0$, there is still acceleration. In this case the velocity vector field does not change with time. But, there can still be a spatial variation which gives rise to a stationary acceleration field and material velocity still changes in space, although the velocity is a constant vector in each point.

Exercise: 3. *Assume 2-D space. Let the temperature field T be given by*

$$T(\vec{r}, t) = T_0 \frac{1}{r} \exp(-t) \qquad t > 0, r > 0$$

The temperature field belongs to a flow field given by $\vec{v}(\vec{r}, t) = \frac{1}{r^2} \exp(-t)\vec{r}$.
a) Determine the divergence of the velocity field.
b) Compute the material derivative of T at any position and time.
Bonus: Compute the particle acceleration.

18.1.4 The material derivative of a material volume integral

Let V be a material volume, i.e. a volume that encompasses for all t the **same flow particles**. This volume is following the flow, possibly being deformed, while there is no material exchange with the region outside V. Let T be again a scalar function of the space and time coordinates. The material derivative of the (material) volume integral of T is (See Appendix T.0.4):

$$\frac{D}{Dt}\left[\int_{V(t)} T(\vec{r}, t)dV\right] = \int_{V(t)} \frac{DT}{Dt} + T\vec{\nabla} \cdot \vec{v} dV = \int_{V(t)} \frac{DT}{Dt} + T\frac{\partial v_k}{\partial x_k} dV \qquad (18.9)$$

> **Exercise: 4.** *Derive from* (18.9) *the alternative formula*
>
> $$\frac{D}{Dt}\left[\int_{V(t)} T(\vec{r},t)dV\right] = \int_{V(t)} \frac{\partial T}{\partial t} + \frac{\partial T v_k}{\partial x_k}dV$$

18.1.5 Diffusion processes

Diffusion processes are in many cases described by (empirical) laws of the form $\vec{a} = -\boldsymbol{D}\cdot\vec{\nabla}H$ where $\vec{a}$ is a vector and $\boldsymbol{D}$ is the (anisotropic) diffusion (coefficient) tensor, and H a scalar field. Examples are Fourier's (isotropic) heat flow vector $\vec{q} = -k\vec{\nabla}T$ where k is thermal conductivity and T temperature, or the isotropic diffusion of matter (atoms) given by the mass density flow vector $\vec{J} = -D\vec{\nabla}c$ where D is the diffusion coefficient and c the concentration of the substance.

18.2 The basic equations of continuum mechanics

> 🇳🇱 De algemene bewegingsvergelijking en de continuiteitsvergelijking vormen de basisvergelijkingen van de continuummechanica. Elke deformeerbare stof voldoet eraan.
>
> | | | |
> |---|---|---|
> | The fluid | $\leftrightarrow$ | de vloiestof |
> | The continuity equation | $\leftrightarrow$ | de continuiteitsvergelijking |
> | The general equation of motion | $\leftrightarrow$ | de algemene bewegingsvergelijking |
> | An incompressible material | $\leftrightarrow$ | en onsamendrukbare stof |
> | Strain rate | $\leftrightarrow$ | de deformatiesnelheid |
> | stress(tensor) | $\leftrightarrow$ | de spanning(stensor) |
> | velocity gradient | $\leftrightarrow$ | |
> | rotation rate | $\leftrightarrow$ | |
> | pressure | $\leftrightarrow$ | de druk |
> | The constitutive equation | $\leftrightarrow$ | de constitutievergelijking |

18.2.1 The continuity equation

The continuity equation is also called mass conservation equation or Reynold's transport theorem. Consider an arbitrary material volume V within a continuum. By definition of a material volume, the mass it contains is conserved, M=constant, or $DM/Dt = 0$. The mass is given by

$$M = \int_{V(t)} \rho(\vec{r},t)dV$$

Applying (18.9) to $DM/Dt = 0$ we find (for arbitrary V) the continuity equation as a local expression of mass conservation:

$$\frac{D\rho}{Dt} + \rho\frac{\partial \mathbf{v}_k}{\partial x_k} = 0 \tag{18.10}$$

or, when substituting the material derivative:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho\mathbf{v}_k)}{\partial x_k} = 0 \qquad \Bigg| \qquad \frac{\partial \rho}{\partial t} + \vec{\nabla}\cdot(\rho\vec{\mathbf{v}}) = 0 \tag{18.11}$$

In incompressible fluids the density cannot change: $D\rho/Dt = 0$. Consequently, mass conservation requires that the divergence of the velocity is 0, i.e.

$$\frac{\partial \mathbf{v}_k}{\partial x_k} = 0 \qquad \Bigg| \qquad \vec{\nabla}\cdot\vec{\mathbf{v}} = 0 \tag{18.12}$$

Note that an equation like (18.10) can also be derived for any other quantity that is conserved by a material volume.

Exercise: 5. *Prove that in a flowing medium with density ρ the following relation holds*

$$\frac{D}{Dt}\left[\int_{V(t)} \rho T dV\right] = \int_{V(t)} \rho \frac{DT}{Dt} dV \qquad (18.13)$$

for any differentiable scalar function T and material volume V. This formula will be frequently used.

Exercise: 6. *a) Prove that in an incompressible fluid the cubic-meter content of a material volume does not change (hence, only the shape of boundary of the material volume is allowed to change).*
b) Prove that in an incompressible fluid the boundary S of a material volume obeys the following integral $\int_S v_k n_k dS = 0$. (Interpret this integral)

Exercise: 7. *Let V be an imaginary volume fixed in space. Derive the alternative mass conservation law:*

$$\int_V \frac{\partial \rho}{\partial t} dV = -\int_S \vec{J} \cdot \vec{n} dS$$

where $\vec{J} = \rho \vec{v}$ is the mass-density flow. (Interpret this equation).

Exercise: 8. *We wish to describe the transport of a polluting substance X carried by a fluid flow. We assume the substance is chemically non-reactive (passive) and dissolved in the fluid. The spatial distribution of X is given by the concentration function $c(\vec{r}, t)$ [kg m^{-3}]. Pollutant is also being produced/destroyed according to the function $H(\vec{r}, t)$ [$\text{kg s}^{-1} \text{m}^{-3}$].*
a) Assume for the moment that mass diffusion of X can be ignored. Derive a conservation law in the form of a differential equation for the concentration function $c(\vec{r}, t)$. [Hint: start with computing the total mass of X (integral form) contained in a material volume V. Next consider the (material) time derivative of this integral. Is equal to what?]
b) Now assume that mass diffusion of the pollutant is important. This implies material diffusion (not controlled by the fluid flow) across the boundary S of the control volume V. Assume that the mass flow density vector $\vec{J}$ [$\text{kg s}^{-1} \text{m}^{-2}$] is given by $\vec{J} = -D\vec{\nabla} c$ where D is the diffusion coefficient. Extend the answer obtained at a) for this situation.

18.2.2 The general equation of motion (momentum equation)

The second law of Newton postulates that the sum of applied forces equals the rate of change of the linear momentum of a particle with mass m:

$$\sum_i \vec{F}_i = \frac{d\vec{p}}{dt}, \qquad \vec{p} = m\vec{v}.$$

To arrive at a similar postulate for continuum mechanics, the total linear momentum of an arbitrary material volume is defined as: $\int_{V(t)} \rho \vec{v} dV$.

Newton's second law leads to the following postulate of continuum mechanics:

$$\frac{D}{Dt}\int_{V(t)} \rho \vec{v} dV = \int_V \rho \vec{g} dV + \int_S \vec{t}^{\vec{n}} dS$$

913

Using (18.13) to evaluate the left side of this equation for each velocity component and applying the derivation following equation (18.3) to the right side we find (as V is arbitrarily chosen):

$$\rho \frac{D\mathbf{v}_i}{Dt} = \frac{\partial \sigma_{ij}}{\partial j} + \rho g_i \qquad \bigg| \qquad \rho \frac{D\vec{\mathbf{v}}}{Dt} = \vec{\nabla} \cdot \boldsymbol{\sigma} + \rho \vec{g} \qquad (18.14)$$

which is the **general equation of motion**. Recall that

$$\frac{D\vec{\mathbf{v}}}{Dt} = \frac{\partial \vec{\mathbf{v}}}{\partial t} + \vec{\mathbf{v}} \cdot \vec{\nabla} \vec{\mathbf{v}}$$

is the material derivative of velocity which renders (18.14) to be a non-linear equation in the unknown velocity field.

18.2.3   Velocity gradient, strain rate, and rotation rate

The velocity gradient tensor is $\vec{\nabla} \vec{\mathbf{v}} = \partial \mathbf{v}_j / \partial x_i$ and can be separated in a symmetric part, the strain rate tensor

$$\dot{\varepsilon}_{ij} = \frac{1}{2} \left(\frac{\partial \mathbf{v}_j}{\partial x_i} + \frac{\partial \mathbf{v}_i}{\partial x_j} \right) \qquad \bigg| \qquad \dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}}) = \frac{1}{2} \left(\vec{\nabla} \vec{\mathbf{v}} + (\vec{\nabla} \vec{\mathbf{v}})^T \right)$$

and in an anti-symmetric part called the rotation rate tensor or spin-rate tensor

$$\dot{\boldsymbol{\omega}}(\vec{\mathbf{v}}) = \frac{1}{2} \left(\vec{\nabla} \vec{\mathbf{v}} - (\vec{\nabla} \vec{\mathbf{v}})^T \right)$$

Note that $\dot{\omega}_{11} = \dot{\omega}_{22} = \dot{\omega}_{33} = 0$. The strain rate tensor is associated with the rate of deformation (rates of relative length and volume changes and shear) while the spin rate tensor describes an increment of uniform rotation in a continuum (i.e. without internal deformation). Note that $\vec{\nabla} \cdot \vec{\mathbf{v}} = \partial \mathbf{v}_k / \partial x_k = \dot{\varepsilon}_{kk}$ gives the rate of relative volume change during deformation.

Exercise: 9. *Make a detailed derivation of equation (18.14).*

Exercise: 10. *Assume a velocity field $\vec{\mathbf{v}} = \vec{\omega} \times \vec{r}$ with $\vec{\omega} = [0, 0, \Omega f(x_1, x_2)]^T$ and $\vec{r} = (x_1, x_2, 0)^T$ where f is an unknown function and Ω is a constant angular speed (radians/sec). Compute the velocity gradient field, the strain rate field and the rotation rate field. Determine a function f that leads to incompressible flow and a function f that leads to flow without shear strain rate.*

18.2.4   Pressure and stress

Similar to traction, the total stress $\boldsymbol{\sigma}(\vec{r}, t)$ is usually separated in the **thermodynamic pressure** p and the rheological (mechanical) stress $\boldsymbol{\pi}(\vec{r}, t)$ (also commonly called viscous stress tensor):

$$\sigma_{ij} = -p\delta_{ij} + \pi_{ij} \qquad \bigg| \qquad \boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\pi} \qquad (18.15)$$

Note: $\boldsymbol{\pi}$ is *not* the deviatoric stress. Keep reading.

In absence of deforming stresses $\boldsymbol{\sigma} = -p\mathbf{1}$ and in the equilibrium state (0 inertial force), Eq. (18.14) reduces to the hydrostatic equation

$$0 = -\frac{\partial p}{\partial x_i} + \rho g_i \qquad \bigg| \qquad \vec{\nabla} p = \rho \vec{g} \qquad (18.16)$$

relating the pressure to the gravitational acceleration.

Exercise: 11. *Let C be a line contour in a continuum, which starts at point A and ends at point B. The continuum is in hydrostatic equilibrium.*
a) Show that the pressure difference between B and A equals $p(B) - p(A) = \int_C \rho \vec{g} \cdot d\vec{r}$ where $d\vec{r}$ is a line element of C.
b) Assume that the continuum is a spherically symmetric body: Show that the pressure difference between B and A is: $p(r_B) - p(r_A) = -\int_{r_A}^{r_B} \rho g(r) dr$ where quantities only depend on the radius.
c) Show that the density field ρ should satisfy $\vec{\nabla}\rho // \vec{\nabla}\Phi$ where Φ is the gravitational energy potential implicitly defined by $\vec{g} = -\vec{\nabla}\Phi$, i.e. the gradient of rho and the gradient of Φ are parallel. [Hint: take the curl of the hydrostatic equation]
d) Show that surfaces of constant pressure, density and gravitational potential coincide.

The average pressure is defined as $\bar{p} = -\frac{1}{3}\sigma_{kk}$. Using (18.15) we find

$$p - \bar{p} = \frac{1}{3}\pi_{kk} \tag{18.17}$$

which demonstrates that local thermodynamic pressure p can be perturbed by the isotropic part of mechanical stress. In fluids this situation can occur in locations of local convergence or divergence of flow (see below).

The deviatoric stress $\boldsymbol{\tau}$ is defined as

$$\tau_{ij} = \sigma'_{ij} = \sigma_{ij} - \frac{1}{3}\sigma_{kk}\delta_{ij} \qquad \Big| \qquad \boldsymbol{\tau} = \boldsymbol{\sigma}' = \boldsymbol{\sigma} - \frac{1}{3}Tr[\boldsymbol{\sigma}]\mathbf{1} \tag{18.18}$$

The deviatoric stress describes the state of stress relative to the ambient (average) pressure. Substituting (18.15) we have

$$\sigma'_{ij} = -(p - \bar{p})\delta_{ij} + \pi_{ij} \tag{18.19}$$

Notice that $\boldsymbol{\tau} = \boldsymbol{\sigma}' = \boldsymbol{\pi}$ when $p = \bar{p}$, i.e. when the mechanical stress does not change the pressure.

Exercise: 12. *Consider a two-dimensional state of stress.*
a) Prove that the principal values of the deviatoric stress are equal in magnitude but opposite in sign by computing the eigen values.
b) As we will see later in the course, the trace of a tensor is invariant (i.e. unchanged) by rotation. Use this information to prove the previous relationship again without calculations.
c) Demonstrate that only shear traction exists on planes bisecting the principal axis of deviatoric stress

---

Exercise: 13. *Demonstrate that the principal deviatoric stresses σ'_i relate to the principal stresses σ_i as: $\sigma'_i = \sigma_i + \bar{p}$*

18.3   Constitutive equations

The mass conservation equation (18.10) and the momentum equation (18.14) constitute 4 equations in 10 unknowns $\rho, \vec{v}, \boldsymbol{\sigma}$ (remember: the stress tensor is symmetric so there are only 6 independent terms out of the 9 it contains). Later we will add the energy equation but this also adds the temperature T as additional unknown. We require knowledge of at least 6 additional independent equations. These are provided for a particular material by specifying the relation between internal kinematics and stress

and are called constitutive equations. For real fluids (i.e. fluids that cannot maintain shear stresses) the constitutive relation involves the viscosity as a material parameter. For solids that can deform brittle, elastic, or exhibit fluid behavior, the constitutive equation(s) will in general involve several material parameters. When stress in a solid material exceeds the elastic strength (a stress limit) the material can either break (deform brittle) or enter a regime of so-called ductile behavior where atoms leave their lattice position and occupy new positions elsewhere. Ductile behavior is accommodated by atomic diffusion processes and by dislocation processes (dislocations are geometric disturbances in a crystalline lattice at which fewer atomic bonds exists and which are thus mechanical weakness zones where applied stress will do its work first). Grain boundary processes are also important agents of deformation but are basically determined by diffusion and dislocation processes. Finding relations between stress and strain rate as a function of material properties, temperature, and pressure is the subject of **rheology**.

The thermodynamic pressure p is taken to be independent of mechanical deformation. The thermodynamic pressure gives the state of stress in a static medium which may have a uniform velocity (either linear, angular of both), i.e. $\dot{\boldsymbol{\varepsilon}} = \mathbf{0}$. Therefore, constitutive equations take the general form

$$\boldsymbol{\sigma}(\dot{\boldsymbol{\varepsilon}}) = -p\mathbf{1} + \boldsymbol{\pi}(\dot{\boldsymbol{\varepsilon}}) \tag{18.20}$$

with $\boldsymbol{\pi}(\mathbf{0}) = \mathbf{0}$, showing that stress and strain rate can be interdependent.

18.3.1   Linear rheology

The most general form of linear rheology leading to a linear viscous fluid (also called a Newtonian fluid) is $\pi_{ij} = C_{ijkl}\dot{\varepsilon}_{kl}$, where C_{ijkl} is the anisotropic viscosity tensor. This relation breaks down to the following law for a purely isotropic linearly viscous fluid:

$$\pi_{ij} = \lambda\dot{\varepsilon}_{kk}\delta_{ij} + 2\eta\dot{\varepsilon}_{ij} \qquad\qquad \boldsymbol{\pi} = \lambda Tr[\dot{\boldsymbol{\varepsilon}}]\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}} \tag{18.21}$$

where η is the dynamic viscosity [Pa s] and λ is a viscosity parameter without a specific name[3].

The mechanical, or rheological, stress π_{ij} quantifies the internal friction of the flow. By computing the trace of the rheological stress we find, using (18.17),

$$p - \overline{p} = \left(\lambda + \frac{2}{3}\eta\right)\dot{\varepsilon}_{kk} = \xi\vec{\nabla}\cdot\vec{v} \tag{18.22}$$

where

$$\xi = \lambda + \frac{2}{3}\eta$$

is the **bulk-viscosity**. Buresti [190] (2015) states: "We may then interpret $\xi\vec{\nabla}\cdot\vec{v}$ as the difference between the thermodynamic pressure and the opposite of the average of the normal stresses acting on any three orthogonal planes passing through a point in the fluid, which is usually referred to as the mechanical pressure. This difference is generally considered to be due to the time lag with which the thermodynamic equilibrium condition is reached in a motion that implies an isotropic dilatation of a fluid element."

Relation (18.22) demonstrates clearly that for a Newtonian fluid the difference between thermodynamic pressure and average pressure is flow induced, i.e. by either local convergence or divergence of flow (i.e. $\vec{\nabla}\cdot\vec{v} \neq 0$).

Remark. *For incompressible fluids $\vec{\nabla}\cdot\vec{v} = 0$. Then, according to (18.22): $p = \overline{p}$. Furthermore, (18.19) leads to $\sigma'_{ij} = \tau_{ij} = \pi_{ij}$.*

---

[3]It is often called the 'second viscosity coefficient'.

Remark. *In the case that the bulk viscosity ξ is assumed 0 (**Stokes condition** - or Stokes hypothesis), we have $p = \overline{p}$ independent of (in)compressibility. Evidently: $p - \overline{p} \leftrightarrow \xi = 0$ or $\vec{\nabla} \cdot \vec{v} = 0$. In geodynamics the Stokes hypothesis is assumed so that in practice we never distinguish $\overline{p}$ from p and in this case $\boldsymbol{\pi} = \boldsymbol{\tau}$ so that the equations are formulated as a function of p and τ (also in the compressible case!). Buresti [190] (2015) states: "It implies that the thermodynamic pressure coincides with the mechanical pressure and characterizes the isotropic part of the complete stress tensor; furthermore, the viscous stress tensor becomes a purely deviatoric tensor and corresponds to the deviatoric part of $\boldsymbol{\sigma}$. In other words, assuming the validity of this hypothesis is equivalent to state that isotropic dilatations of an elementary volume of fluid do not produce viscous stresses."*

Note that in some books/publications one can find a different formulation of the Stokes hypothesis. For example in Carey and Oden [224] (1986) we find that "Stokes' hypothesis is that $\boldsymbol{\pi}$ is a function only of the deformation rate tensor $\dot{\boldsymbol{\varepsilon}}$".

In terms of bulk viscosity and dynamic viscosity Eq. (18.21) reads

$$\pi_{ij} = \left(\xi - \frac{2}{3}\eta\right)\dot{\varepsilon}_{kk}\delta_{ij} + 2\eta\dot{\varepsilon}_{ij}. \qquad\qquad \boldsymbol{\pi} = \left(\xi - \frac{2}{3}\eta\right)\dot{\varepsilon}_{kk}\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}. \qquad (18.23)$$

By using the equation of deviatoric strain rate,

$$\dot{\varepsilon}'_{ij} = \dot{\varepsilon}_{ij} - \frac{1}{3}\dot{\varepsilon}_{kk}\delta_{ij} \qquad\qquad \dot{\boldsymbol{\varepsilon}}' = \dot{\boldsymbol{\varepsilon}} - \frac{1}{3}\dot{\varepsilon}_{kk}\mathbf{1} \qquad (18.24)$$

we obtain as alternative expression for Eq. (18.21)

$$\pi_{ij} = \xi\dot{\varepsilon}_{kk}\delta_{ij} + 2\eta\dot{\varepsilon}'_{ij} \qquad\qquad \boldsymbol{\pi} = \xi\dot{\varepsilon}_{kk}\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}' \qquad (18.25)$$

which explicitly shows the role of bulk viscosity and dynamic viscosity. In particular, for an incompressible fluid we have

$$\tau_{ij} = \pi_{ij} = 2\eta\dot{\varepsilon}'_{ij} = 2\eta\dot{\varepsilon}_{ij} \qquad\qquad \boldsymbol{\tau} = \boldsymbol{\pi} = 2\eta\dot{\boldsymbol{\varepsilon}}' = 2\eta\dot{\boldsymbol{\varepsilon}} \qquad (18.26)$$

From Eq. (18.25) we can write

$$\boldsymbol{\sigma} = (-p + \xi\vec{\nabla} \cdot \mathbf{v})\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}'$$

Exercise: 14. *Prove that for a Newtonian fluid $\sigma'_{ij} = 2\eta\dot{\varepsilon}'_{ij}$ where the prime denotes the deviators of stress and strain rate. (Interpret this general equation)*

18.3.2 Non-linear rheology

Microphysical processes in solids like atomic diffusion and dislocation motion lead to permanent deformation (macroscopic flow). Pure diffusion, either along grain boundaries or through the bulk of a grain, is a prime example of a Newtonian deformation mechanism. Dislocation glide (low-temperature creep or exponential creep) and diffusion assisted dislocation climb (power law creep) are examples of deformation mechanisms with a nonlinear relation between stress and strain rate. Microphysical considerations (theory) combined with lab experiments (practice) lead to the following (simplified) constitutive equation relating strain rate to rheological stress:

$$\dot{\varepsilon}_{ij} = A^{-1}\tau^{n-1}\tau_{ij} \qquad (18.27)$$

where $n \geq 1$ is the stress exponent, A is a material parameter and $\tau = (\frac{1}{2}\tau_{ij}\tau_{ij})^{1/2}$ is the effective stress, i.e. the second invariant of rheological stress (i.e. a scalar stress quantity which value is independent of coordinate frame). In analogy with linear viscosity in an incompressible fluid we define the viscosity function

$$\eta = \frac{\tau_{ij}}{2\dot{\varepsilon}_{ij}}. \tag{18.28}$$

Combined with (18.27), the viscosity function of power law rheology reads:

$$\eta(\tau) = \frac{1}{2}A\tau^{1-n}. \tag{18.29}$$

Notice that the viscosity decreases as stress increases. The stress (internal friction) is due to body forces and/or applied tractions (cf. Eq. (18.14)). In regions of high internal stress, the viscosity decreases in a power law rheology which causes increasing strain rates (according to $\tau_{ij} = 2\eta\dot{\varepsilon}_{ij}$) leading to a localization of deformation.

Exercise: 15. *Derive the following relations:*
a) $\tau = (A\dot{\varepsilon})^{1/n}$
b) $\tau_{ij} = A^{1/n}\dot{\varepsilon}^{-1+1/n}\dot{\varepsilon}_{ij}$
c) $\eta(\dot{\varepsilon}) = \frac{1}{2}A^{1/n}\dot{\varepsilon}^{-1+1/n}$
where $\dot{\varepsilon}$ is the effective strain rate.

In the upper mantle $n \simeq 3 - 5$ while in the lower mantle $n \simeq 1 - 3$ (but in both cases not known for sure!). In theory, if $\dot{\varepsilon} \to 0$ then $\eta \to \infty$ which leads to stagnating flow. In practice, however, other (competing) deformation mechanism will provide higher strain rates keeping the viscosity finite. The total strain rate is the sum of strain rate contributions from the different mechanisms active under the same rheological stress:

$$\dot{\varepsilon}_{ij} = \sum_k \dot{\varepsilon}_{ij}^{(k)} = \frac{1}{2}\sum_k (\eta^{(k)})^{-1}\tau_{ij} = \frac{1}{2}\tau_{ij}\sum_k (\eta^{(k)})^{-1} = \frac{\tau_{ij}}{2\eta_{eff}}$$

where the effective viscosity is

$$\eta_{eff} = \sum_k \frac{1}{\eta^{(k)}} \tag{18.30}$$

A constitutive equation, more detailed than (18.27), encompassing both power law creep and pure diffusion creep is (see for example Karato and Wu [714] (1993)):

$$\dot{\varepsilon}_{ij} = A\left(\frac{b}{d}\right)^m \exp\left(-\frac{Q+pV}{RT}\right)\left(\frac{\tau}{\mu}\right)^{n-1}\tau_{ij} \tag{18.31}$$

with

- b: length of Burgers (dislocation) vector ($\sim 5 \cdot 10^{-10}$m)

- d: grain size (0.001-0.1 m)

- Q: activation energy (related to atomic bonding)

- V: activation volume (associated with atomic diffusion)

- R: Universal gas constant 8.31444 J/(mol K)

- p, T: pressure and temperature

- μ: elastic shear modulus (only used to scale stress) (80 GPa)

The combination of $n = 1$ and $m = 2$ or 3, gives a Newtonian fluid resulting from pure atomic diffusion. $n > 1$ and $m = 0$ relates to various dislocation (power law) creep mechanisms.

Furthermore, feedback relations can exist between grain growth (due to dynamic crystallization) and stress, e.g.

$$d = Kb \left(\frac{\tau}{\mu} \right)^{-q} \qquad K \sim 19 \tag{18.32}$$

The above account of deformation laws (constitutive equations) is by no means exhaustive and only presented to give examples of nonlinear relations between stress and strain rate involving a viscosity function and, in practice, leading to the notion of effective viscosity derived from a superposition of competing deformation mechanisms.

<div style="background:orange">ADD effective viscosity formula</div>

> **Exercise: 16.** *a) What is the relative change in effective viscosity assuming both pressure and tempoerature are increased by 10% ?*
> *b) Same question when only temperature is increased by 10% ?*

18.4 The Navier-Stokes equation

Recalling the general equation of motion (18.14)

$$\rho \frac{D\mathbf{v}_i}{Dt} = \rho g_i + \frac{\partial \sigma_{ij}}{\partial x_j} \qquad \Big| \qquad \rho \frac{D\vec{\mathbf{v}}}{Dt} = \rho \vec{g} + \vec{\nabla} \cdot \boldsymbol{\sigma}$$

and the separation of the total stress in the thermodynamic pressure and mechanical stress (18.15)

$$\sigma_{ij} = -p\delta_{ij} + \pi_{ij}, \qquad \Big| \qquad \boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\pi} \tag{18.33}$$

we find after substitution:

$$\rho \frac{D\mathbf{v}_i}{Dt} = \rho g_i - \frac{\partial p}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} \qquad \Big| \qquad \rho \frac{D\vec{\mathbf{v}}}{Dt} = \rho \vec{g} - \vec{\nabla} p + \vec{\nabla} \cdot \boldsymbol{\pi} \tag{18.34}$$

Adopting the constitutive equation (18.23) for linear rheology

$$\pi_{ij} = \left(\xi - \frac{2}{3}\eta \right) \dot{\varepsilon}_{kk}\delta_{ij} + 2\eta\dot{\varepsilon}_{ij} \qquad \Big| \qquad \boldsymbol{\pi} = \left(\xi - \frac{2}{3}\eta \right) \dot{\varepsilon}_{kk}\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}} \tag{18.35}$$

and assuming that viscosities do not depend on the spatial coordinates we find that

$$\rho \frac{D\mathbf{v}_i}{Dt} = \rho g_i - \frac{\partial p}{\partial x_i} + \eta\vec{\nabla}^2\mathbf{v}_i + \left(\xi + \frac{1}{3}\eta \right) \frac{\partial}{\partial x_i}\vec{\nabla}\cdot\vec{\mathbf{v}} \quad \Big| \quad \rho \frac{D\vec{\mathbf{v}}}{Dt} = \rho \vec{g} - \vec{\nabla} p + \eta\vec{\nabla}^2\vec{\mathbf{v}} + \left(\xi + \frac{1}{3}\eta \right) \vec{\nabla}(\vec{\nabla}\cdot\vec{\mathbf{v}})$$
$$\tag{18.36}$$

This is the **Navier-Stokes equation** for a fluid with constant viscosities.

<div style="background:orange">check this equation!</div>

> **Exercise: 17.** *Derive* (18.36)

18.5 Density perturbations as a driving force for mantle convection

The general equation of motion obtained previously

$$\rho \frac{D\mathbf{v}_i}{Dt} = -\frac{\partial p}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} + \rho g_i \qquad \Bigg| \qquad \rho \frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla} p + \vec{\nabla} \cdot \boldsymbol{\pi} + \rho \vec{g} \qquad (18.37)$$

can be rewritten in terms of an equation relative to the hydrostatic reference state of the Earth's mantle. We define the hydrostatic reference state as the undeformed state in which no density perturbations exist. The hydrostatic, or static, pressure is $p_0(r)$ and the density field in the hydrostatic state is $\rho_0(r)$. For the acceleration of gravity in the hydrostatic state we take $g_i^0(r)$. The hydrostatic quantities obey the equilibrium equation:

$$0 = -\frac{\partial p_0}{\partial x_i} + \rho_0 g_i^0. \qquad \Bigg| \qquad \vec{0} = -\vec{\nabla} p_0 + \rho_0 \vec{g}^{\,0} \qquad (18.38)$$

We separate quantities in the dynamic state in a hydrostatic contribution and a perturbation:

$$
\begin{aligned}
p(\vec{r}) &= p_0(\vec{r}) + \tilde{p}(\vec{r}) \\
\rho(\vec{r}) &= \rho_0(\vec{r}) + \tilde{\rho}(\vec{r}) \\
\vec{g}(\vec{r}) &= \vec{g}_0(\vec{r}) + \vec{\tilde{g}}(\vec{r})
\end{aligned}
$$

where $\tilde{p}(\vec{r})$ is the so-called dynamic pressure. Note that the mechanical stress $\boldsymbol{\pi}$ is by itself a perturbation with respect to the hydrostatic state. After substitution in the equation of motion the hydrostatic terms cancel and we find:

$$(\rho_0 + \tilde{\rho}) \frac{D\mathbf{v}_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} + \tilde{\rho} g_i^0 + \tilde{\rho} \tilde{g}_i + \rho_0 \tilde{g}_i \qquad (18.39)$$

The term $\tilde{\rho} D\mathbf{v}_i/Dt$ and the the last two terms can often be neglected as a small second order perturbation leading to the perturbation equation

$$\rho_0 \frac{D\mathbf{v}_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} + \tilde{\rho} g_i^0 \qquad \Bigg| \qquad \rho_0 \frac{D\vec{\mathbf{v}}}{Dt} = -\vec{\nabla} \tilde{p} + \vec{\nabla} \cdot \boldsymbol{\pi} + \tilde{\rho} \vec{g} \qquad (18.40)$$

This equation shows explicitly that density perturbations and not the total density field are the driving forces for mantle convection.

Gravity inside the Earth relates to the density field according to the differential (Poisson) equation:

$$\vec{\nabla} \cdot \vec{g} = -\vec{\nabla}^2 U = -4\pi \mathcal{G} \rho \tag{18.41}$$

where $\mathcal{G}$ is the universal constant of gravity. With $\vec{g} = -\vec{\nabla}U$, we obtain the **Poisson equation**:

$$\vec{\nabla}^2 U = 4\pi \mathcal{G} \rho \tag{18.42}$$

or, for the dynamic quantities deviating from the hydrostatic state

$$\vec{\nabla}^2 \tilde{U} = 4\pi \mathcal{G} \tilde{\rho} \tag{18.43}$$

Exercise: 19. *Assume a spherically symmetric non-rotating Earth in hydrostatic equilibrium. In spherical coordinates the divergence of a vector field $\vec{a}(r)$, which only depends on the radius r is*

$$\vec{\nabla} \cdot \vec{a} = \frac{1}{r^2} \frac{d}{dr}(r^2 a_r)$$

a) Prove that the acceleration of gravity at radius r only depends on the mass contained in the sphere of radius R.

b) Assume that the mass of the Earth's core is M_c. Assume a linear density profile for the crust and mantle and determine the acceleration of gravity as a function of the radius in the mantle.

18.6 Two-dimensional formulation for incompressible fluids: the stream function approach

Although 2-D formulations of flow problems may seem restrictive at first glance, it is useful for many applications in which variation in the omitted dimension are small. For instance, slab subduction of a laterally long subduction zone can be modeled in 2-D perpendicular to the strike of a subduction zone. Apart from applications, a full 3-D approach is not always necessary in order to get insight into fundamental problems like the onset of convection, or studies of the influence on the flow pattern of certain model parameters (e.g. viscosity function, internal heat production), or of specific model attributes (e.g. phase transitions, boundary layers). Of course, before the age of high- speed computers, flow calculations where done mostly analytically which is easier in 2-D than in 3-D.

In 2-D we have the 4 unknowns: v_x, v_y, ρ and p[4]. By assuming incompressibility we can satisfy this equation trivially by defining the stream function $\psi(x, y)$ as follows:

$$
\begin{aligned}
v_x &= \frac{\partial \psi}{\partial y} \\
v_y &= -\frac{\partial \psi}{\partial x}
\end{aligned}
\tag{18.44}
$$

Please check that we have indeed $\vec{\nabla} \cdot \vec{v} = 0$. This definition does not impose any restriction on the velocity field or the stream function (apart from being a differentiable function).

Furthermore, assume (for now) that the viscosity is constant and that $\rho D\vec{v}/DT \sim 0$ for mantle convection (this will be demonstrated later). Then, the Navier-Stokes equation (18.36) reduces to

---

[4]Here too we assume Stokes hypothesis, so that $\bar{p} = p$ and $\boldsymbol{\pi} = \boldsymbol{\tau}$

$-\vec{\nabla} p + \eta \vec{\nabla}^2 \vec{v} + \rho \vec{g} = \vec{0}$ and then :

$$
\begin{aligned}
0 &= -\frac{\partial p}{\partial x} + \eta \left(\frac{\partial^3 \psi}{\partial y^3} + \frac{\partial^3 \psi}{\partial x^2 \partial y} \right) + \rho g_x \\
0 &= -\frac{\partial p}{\partial y} - \eta \left(\frac{\partial^3 \psi}{\partial x^3} + \frac{\partial^3 \psi}{\partial y^2 \partial x} \right) + \rho g_y
\end{aligned}
\tag{18.45}
$$

The pressure terms in (18.45) can be removed by first differentiating the first line w.r.t. y and the second line w.r.t. x,

$$
\begin{aligned}
0 &= -\frac{\partial^2 p}{\partial x \partial y} + \eta \left(\frac{\partial^4 \psi}{\partial y^4} + \frac{\partial^4 \psi}{\partial x^2 \partial y^2} \right) + \frac{\partial(\rho g_x)}{\partial y} \\
0 &= -\frac{\partial^2 p}{\partial y \partial x} - \eta \left(\frac{\partial^4 \psi}{\partial x^4} + \frac{\partial^4 \psi}{\partial y^2 \partial x^2} \right) + \frac{\partial(\rho g_y)}{\partial x}
\end{aligned}
\tag{18.46}
$$

and next by subtracting the resulting equations, leading to:

$$
0 = \eta \left(\frac{\partial^4 \psi}{\partial x^4} + 2\frac{\partial^4 \psi}{\partial x^2 \partial y^2} + \frac{\partial^4 \psi}{\partial y^4} \right) + \frac{\partial(\rho g_x)}{\partial y} - \frac{\partial(\rho g_y)}{\partial x}
\tag{18.47}
$$

or

$$
\vec{\nabla}^4 \psi = \frac{1}{\eta} \left(-\frac{\partial(\rho g_x)}{\partial y} + \frac{\partial(\rho g_y)}{\partial x} \right)
\tag{18.48}
$$

where

$$
\vec{\nabla}^4 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)
$$

Equation (18.48) is the inhomogeneous bi-harmonic equation.

Exercise: 20. *Show that (18.48) reduces to the homogeneous bi-harmonic equation $\vec{\nabla}^4 \psi = 0$ if density is independent of the spatial coordinates. NB: we assumed $\vec{\nabla} \cdot \vec{v} = 0$ which requires, because of the continuity equation, that density is constant in time.*

Exercise: 21. *Consider an incompressible flow in three dimensions and (implicitly) define the vector potential $\vec{\psi}$ as $\vec{v} = -\vec{\nabla} \times \vec{\psi}$.*
Assume a Newtonian fluid with constant viscosity and assume that $\rho D v_i / Dt \sim 0$.
Demonstrate that

$$
\vec{\nabla} \times \vec{\nabla} \times \vec{\nabla} \times \vec{\nabla} \times \vec{\psi} = -\frac{1}{\eta} \vec{\nabla} \times (\rho \vec{g})
$$

(Use the identity $\vec{\nabla}^2 \vec{v} = \vec{\nabla}(\vec{\nabla} \cdot \vec{v}) - \vec{\nabla} \times \vec{\nabla} \times \vec{v}$ and take the curl of the N-S equation)
Bonus: Demonstrate the well known identity $\vec{\nabla} \cdot \vec{v} = 0$.

18.6.1 Application of the stream function approach: Post-glacial rebound

We consider the restoration of the Earth's surface to equilibrium shape in the aftermath of global de-glaciation. Relatively sudden removal of the huge ice-caps, covering parts of the northern and southern hemisphere, left a depression in the Earth surface which, as a result of horizontal pressure gradients, is being restored to isostatic equilibrium. The uplift resulting from the last glaciation period started some 8000 years ago and is as yet not complete. To model the isostatic rebound we

adopt an approximate analytical approach originally due to Haskell [585] (1935) - see also Mitrovica [931] (1996). This analysis will lead to a first estimate of the viscosity of the mantle.

Assume an infinite 2-D half-space filled with an incompressible Newtonian fluid of constant viscosity and with density independent of spatial coordinates. The y-axis is taken positive downward. The equations to solve are (18.48) for the stream function and next (18.45) for the pressure. As density is constant the right-hand-side of (18.48) is zero. Furthermore, we assume that the horizontal component of gravity is 0 which allows for a straightforward solution of (18.45) when the stream function is known. Because (18.48) is a linear differential equation we assume a spectral approach in which surface deformation is prescribed as a harmonic function with a specific wavelength λ :

$$w_k(x,t) = w_0(t)\cos(kx) \tag{18.49}$$

with the wave number $k = 2\pi/\lambda$.

In this model the surface deflection results from sudden ice-unloading at $t = 0$ and $w_0(t) \to 0$ if $t \to \infty$. Because of the linearity of (18.48) the 'loading' function $w_k(x,t)$ will lead to a stream function of the form $\psi(x,y) = A(y)\cos kx + B(y)\sin kx$ (i.e. harmonic with separation of variables). A more detailed analysis than is given below will show that $A(y) = 0$. For simplicity, we take

$$\psi(x,y) = Y(y)\sin kx \tag{18.50}$$

Substitution in (18.48) leads to

$$\frac{d^4 Y}{dy^4} - 2k^2\frac{d^2 Y}{dy^2} + k^4 Y = 0 \tag{18.51}$$

This is an ordinary differential equation (ODE) with constant coefficients. Substituting $Y(y) = Y_0\exp(my)$ gives

$$m^4 - 2k^2 m^2 + k^4 = (m^2 - k^2)^2 = 0$$

hence $m = \pm k$ with multiplicity 2. The multiplicity requires two additional solutions $y\exp(\pm ky)$. The general solution is then

$$Y(y) = A\exp(-ky) + By\exp(-ky) + C\exp(ky) + Dy\exp(ky) \tag{18.52}$$

To arrive at a specified solution for the problem we have to consider boundary conditions. The first condition is that when $y \to \infty$ then $\psi \to 0$. This guarantees a finite solution at infinity and satisfies the idea that at large depth the velocity field should approach zero. Substitution in (18.52) leads to $C = D = 0$ and to the stream function solution

$$\psi(x,y) = (A + By)\exp(-ky)\sin kx \tag{18.53}$$

The second boundary condition concerns the motion of the surface. We expect that the deflection $w_k(x,t)$ primarily leads to a vertical motion for small wave number k (large wavelength λ) and assume that horizontal motion is negligibly small[5]. The condition $\mathsf{v}_x(x,y) = 0$ for $y = w_k(x,t)$ is only a crude approximation. Still, for Haskell's problem it helps finding an analytical solution. Using (18.44) we find for the velocity components:

$$\mathsf{v}_x = (B - k(A + By))\exp(-ky)\sin kx \tag{18.54}$$
$$\mathsf{v}_y = -k(A + By)\exp(-ky)\cos kx \tag{18.55}$$

Exercise: 22. *Derive equations* (18.54) *and* (18.55).

---

[5]From modern GPS research on surface motions in Scandinavia we know that horizontal surface motions are between 10-20% of the vertical motion

The condition $\nu_x(x, w_k(x, t)) = 0$ leads to $B \simeq kA$ where we used that $kw_k = 2\pi w_k/\lambda \ll 1$ because w_k is on the order of 1 km while λ is on the order of 3000 km. This leads to the following results:

$$\psi(x, y) = A(1 + ky)\exp(-ky)\sin kx \qquad (18.56)$$

$$\nu_x = -Ak^2 y \exp(-ky)\sin kx \qquad (18.57)$$

$$\nu_y = -Ak(1 + ky)\exp(-ky)\cos kx \qquad (18.58)$$

The last boundary condition concerns the zero pressure at the deformed surface $y = w_k(x, t)$. To solve for the pressure the solution (18.56) is first substituted in (18.45) which gives

$$0 = -\frac{\partial p}{\partial x} + 2\eta A k^3 \exp(-ky)\sin kx + \rho g_x \qquad (18.59)$$

$$0 = -\frac{\partial p}{\partial y} + 2\eta A k^3 \exp(-ky)\cos kx + \rho g_y \qquad (18.60)$$

Exercise: 23. *Derive equations (18.59) and (18.60).*

With $g_x \simeq 0$, integration of (18.60) leads to

$$p(x, y) = -2\eta A k^2 \exp(-ky)\cos kx + \rho g_y y + p_0 \qquad (18.61)$$

Exercise: 24. *Derive Eq. (18.61) and justify the assumption $p_0 = 0$.*

Using that $kw_k \ll 1$, the condition $p(x, w_k(x, t)) = 0$ gives for the surface deflection

$$w_k(x, t) = \frac{2A\eta k^2}{\rho g_y}\cos kx \qquad (18.62)$$

This solution explicitly relates the amplitude of the surface deflection to the wave length $\lambda = 2\pi/k$ and the principal model parameters density and viscosity. However, there is still an undetermined coefficient A which must describe the time behavior of the surface deformation (recall that $w_k(x, t) \to 0$ if $t \to \infty$). Time has played no role in the derivation because we have basically solved a time-stationary process. Time stationary problems are characterized by a static velocity field (time-constant velocity vectors at each position). But, the velocity field still describes material flow and time is a parameter if we wish to follow a particle in the flow. Particularly, the surface particles move with the vertical velocity $\nu_y(x, w_k(x, t))$ and thus

$$\frac{\partial w_k}{\partial t} = \nu_y(x, w_k(x, t)) = -Ak(1 + kw_k)\exp(-kw_k)\cos kx \simeq -Ak\cos kx \qquad (18.63)$$

Using (18.62) we can eliminate A and arrive at the differential equation

$$\frac{\partial w_k}{\partial t} = -\frac{\rho g_y}{2\eta k}w_k \qquad (18.64)$$

which has the solution

$$w_k(x, t) = w_k(x, 0)\exp(-t/t_r) \qquad (18.65)$$

where the relaxation time is $t_r = 4\pi\eta/\rho g_y \lambda$.

From geological observations of (e.g. beach uplift, river incisions), data curves of $w(x, t)$ have been obtained which leads to relaxation times of about 4400 yr for wave lengths of about 3000 km. Using values of 3300 kg m^{-3} for density and 10 m s^{-2} for the acceleration of gravity, we find for the viscosity $\eta \simeq 10^{21}$ Pa s, a number that still stands today as an average value for the 600 to 1000 km of the mantle.
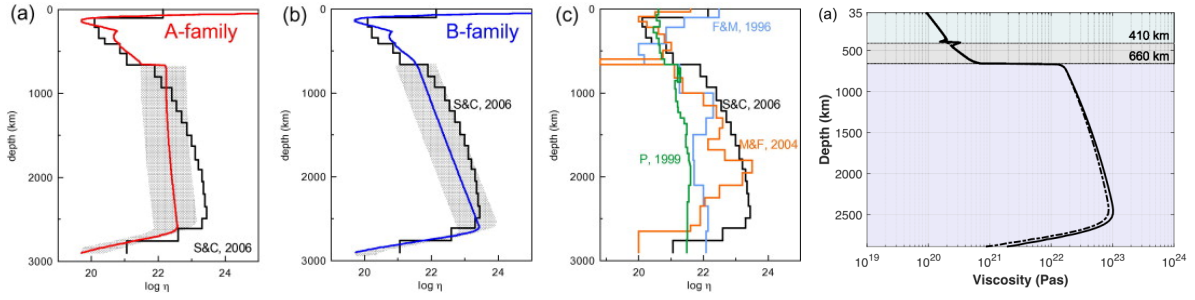
Left: Taken from Ciskova *et al.* [277] (2012); Right: Taken from Neuharth and Mittelstaedt [991] (2023).

Modern modeling concentrates on the complex global problem of Global Isostatic Adjustment (GIA) which also involves the global variation of sea-level which is coupled to ice-sheet formation and melting. Both are functions of time and surface coordinates (topography!). Positive feedback exist between water extraction from the oceans (unloading; attraction of mantle flow) and ice-sheet creation elsewhere (loading; pushing the mantle away). Inverse processes occur during and after ice cap melting. Much data is available on relative sea-level changes, but much less about actual ice sheet (un-)loading histories. Results obtained from different GIA modeling strategies still show disagreement on the detail of viscosity change with depth which also results from the relatively poor sensitivity of surface motions for the detail of viscosity layering. The interested student can find its way in the literature through papers of J. Mitrovica, D. Peltier, and K. Lambeck.

18.7 The energy equation

Mantle convection is driven by density perturbations relative to a hydrostatic state. In part, the density perturbations $\Delta\rho$ are due to thermal perturbations ΔT where the connection is given by an equation of state involving the thermal expansion coefficient α. The full description of mantle convection requires also an equation describing the temperature field of flow. This equation is called the energy equation and involves contributions from adiabatic (de)compression, heat dissipation due to friction in the flow (viscous dissipation), heat conduction and advection, and heat production (including phase changes).

The derivation of the energy equation starts with the first law of thermodynamics which equates the change ΔE in total energy of a system to the work ΔW done by thermo-mechanical processes and the heat input ΔQ to the system resulting from heat flow and heat production (we disregard here the energy contribution from chemical and electro- magnetic processes):

$$\Delta E = \Delta W + \Delta Q \tag{18.66}$$

For processes developing continuously in time we can write (18.66) in terms of the power

$$\frac{DE}{Dt} = \frac{DW}{Dt} + \frac{DQ}{Dt} \tag{18.67}$$

To apply the first law of thermodynamics to the mechanical deformation of a continuum, consider a material volume V with boundary S. The power[6] developed by mechanical forces $\rho\vec{g}$ and the boundary tractions $\vec{t}^n$ is

$$\frac{DW}{Dt} = \int_V \rho g_i \mathbf{v}_i dV + \int_S \mathbf{v}_i t_i^n dS \qquad \Bigg| \qquad \frac{DW}{Dt} = \int_V \rho\vec{g} \cdot \vec{\mathbf{v}} dV + \int_S \vec{\mathbf{v}} \cdot \vec{t}^h dS \tag{18.68}$$

By substituting the Cauchy relation (18.2) (i.e. $t_i^n = \sigma_{ij}n_j$ or $\vec{t}^h = \boldsymbol{\sigma} \cdot \vec{n}$), with $\vec{n}$ the outward pointing normal on S, and next applying the divergence theorem, the surface integral is transformed to a volume integral:

$$\begin{aligned}
\frac{DW}{Dt} &= \int_V \rho\vec{\mathbf{v}} \cdot \vec{g} dV + \int_S \vec{\mathbf{v}} \cdot (\boldsymbol{\sigma} \cdot \vec{n}) dS \\
&= \int_V \rho\vec{\mathbf{v}} \cdot \vec{g} dV + \int_S (\vec{\mathbf{v}} \cdot \boldsymbol{\sigma}) \cdot \vec{n} dS \\
&= \int_V \rho\vec{\mathbf{v}} \cdot \vec{g} dV + \int_V \vec{\nabla} \cdot (\vec{\mathbf{v}} \cdot \boldsymbol{\sigma}) dV \\
&= \int_V \vec{\mathbf{v}} \cdot (\rho\vec{g}) dV + \int_V \vec{\mathbf{v}} \cdot \vec{\nabla} \cdot \boldsymbol{\sigma} dV + \int_V \vec{\nabla}\vec{\mathbf{v}} : \boldsymbol{\sigma} dV
\end{aligned}$$

After rearranging terms we get

$$\frac{DW}{Dt} = \int_V \left[\mathbf{v}_i \left(\rho g_i + \frac{\partial\sigma_{ij}}{\partial x_j} \right) + \sigma_{ij}\frac{\partial\mathbf{v}_i}{\partial x_j} \right] dV \quad \Bigg| \quad \frac{DW}{Dt} = \int_V \left[\vec{\mathbf{v}} \cdot (\rho\vec{g} + \vec{\nabla} \cdot \boldsymbol{\sigma}) + \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} \right] dV \tag{18.69}$$

The general equation of motion (18.14) is used in (18.69) to arrive at the second equality. Next, we

---

[6]`https://en.wikipedia.org/wiki/Power_(physics)`

can apply (18.13) to the first term in the last equality which leads to

$$
\begin{aligned}
\frac{DW}{Dt} &= \int_V \left[\vec{\mathbf{v}} \cdot (\rho \vec{g} + \vec{\nabla} \cdot \boldsymbol{\sigma}) + \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} \right] dV \\
&= \int_V \left[\vec{\mathbf{v}} \cdot \left(\rho \frac{D\vec{\mathbf{v}}}{Dt} \right) + \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} \right] dV \\
&= \int_V \left[\rho \frac{D(\frac{1}{2}\vec{\mathbf{v}} \cdot \vec{\mathbf{v}})}{Dt} + \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} \right] dV \\
&= \frac{D}{Dt} \int_V \rho \frac{1}{2} \vec{\mathbf{v}} \cdot \vec{\mathbf{v}} \, dV + \int_V \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} \, dV \qquad (18.70)
\end{aligned}
$$

or,

$$
\begin{aligned}
\frac{DW}{Dt} &= \int_V \left[\mathbf{v}_i \left(\rho g_i + \frac{\partial \sigma_{ij}}{\partial x_j} \right) + \sigma_{ij} \frac{\partial \mathbf{v}_i}{\partial x_j} \right] dV \\
&= \int_V \left(\rho \mathbf{v}_i \frac{D\mathbf{v}_i}{Dt} + \sigma_{ij} \frac{\partial \mathbf{v}_i}{\partial x_j} \right) dV \\
&= \int_V \left(\frac{1}{2}\rho \frac{D(\mathbf{v}_i \mathbf{v}_i)}{Dt} + \sigma_{ij} \frac{\partial \mathbf{v}_i}{\partial x_j} \right) dV \\
&= \frac{D}{Dt} \int_V \frac{1}{2}\rho \mathbf{v}_i \mathbf{v}_i \, dV + \int_V \sigma_{ij} \frac{\partial \mathbf{v}_i}{\partial x_j} \, dV \qquad (18.71)
\end{aligned}
$$

The first term in (18.71) describes the power input from kinetic energy of the flow and the second, the power input resulting from viscous dissipation.

The term DQ/Dt in (18.67) concerns the heat flow $\vec{q}(\vec{r},t)$ across the boundary S and the heat production $H(\vec{r},t)$ within the volume V. The heat flow $\vec{q}$ has dimensions $\mathrm{J\,s^{-1}\,m^{-2}}$ and caused by thermal gradients in the medium as described by the Fourier law $\vec{q} = -k\vec{\nabla}T$ where k is the thermal conductivity. The heat production H (or consumption as in an endothermic phase change) has units $\mathrm{J\,s^{-1}\,kg^{-1}}$. The total rate of heat input is obtained by integration of the heat production over the volume and by integration of the normal component of heat flow over the surface S

$$
\begin{aligned}
\frac{DQ}{Dt} &= \int_V \rho H dV - \int_S \vec{q} \cdot \vec{n} \, dS \qquad &(18.72) \\
&= \int_V \rho H dV - \int_V \vec{\nabla} \cdot \vec{q} \, dV \qquad &(18.73) \\
&= \int_V \left(\rho H + \vec{\nabla} \cdot (k\vec{\nabla}T) \right) dV \qquad &(18.74)
\end{aligned}
$$

Outward directed heat flow causes a negative contribution to DQ/Dt as heat is flowing out of the volume. Because the normal $\vec{n}$ is defined as outward pointing on S a minus sign is therefore required in front of the heat flow integral in the first equality. In the second equality, the Fourier law for heat conduction is substituted. Application of the divergence theorem leads to the last equality.

The total energy E in (18.67) is now separated into two contributions: the kinetic energy and the internal energy[7] e ($\mathrm{J\,kg^{-1}}$)[8]:

$$
E = \int_V \left(\frac{1}{2}\rho \mathbf{v}_i \mathbf{v}_i + \rho e \right) dV \qquad \bigg| \qquad E = \int_V \left(\frac{1}{2}\rho \vec{\mathbf{v}} \cdot \vec{\mathbf{v}} + \rho e \right) dV \qquad (18.75)
$$

---

[7] https://en.wikipedia.org/wiki/Internal_energy
[8] $[e] = L^2 T^{-2}$

Taking the time derivative leads to

$$
\begin{aligned}
\frac{DE}{Dt} &= \frac{D}{Dt}\left(\int_V \frac{1}{2}\rho\vec{\mathbf{v}}\cdot\vec{\mathbf{v}}dV\right) + \frac{D}{Dt}\left(\int_V \rho e\, dV\right) \\
&= \frac{D}{Dt}\left(\int_V \frac{1}{2}\rho\vec{\mathbf{v}}\cdot\vec{\mathbf{v}}dV\right) + \left(\int_V \rho\frac{De}{Dt}dV\right)
\end{aligned}
\tag{18.76}
$$

where we applied equation (18.13) to rewrite the last integral.

We are now ready to create the energy balance of Eq. (18.67): Combining equations (18.76), (18.71), and (18.74) results in an integral equation over the material volume V in which the kinetic energy terms cancel. As V is chosen arbitrary we find the **energy (or heat) equation**

$$
\rho\frac{De}{Dt} = \sigma_{ij}\frac{\partial \mathbf{v}_i}{\partial x_j} + \frac{\partial}{\partial x_j}\left(k\frac{\partial T}{\partial x_j}\right) + \rho H \qquad\Bigg|\qquad \rho\frac{De}{Dt} = \boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\cdot(k\vec{\nabla}T) + \rho H
\tag{18.77}
$$

which states that the rate of change of internal energy equals the sum of viscous heat dissipation, thermal conduction, and heat production.

Remark. *Looking at units: We have the internal energy e given in J/kg, i.e. $[e] = ML^2T^{-2}/M = L^2T^{-2}$. Then $[\rho De/Dt] = ML^{-3}(L^2T^{-2})T^{-1} = ML^{-1}T^{-3}$, $[\boldsymbol{\sigma} : \vec{\nabla}\vec{\mathbf{v}}] = ML^{-1}T^{-2}L^{-1}LT^{-1} = ML^{-1}T^{-3}$, $[\vec{\nabla}\cdot(k\vec{\nabla}T)] = L^{-1}LMT^{-3}\theta^{-1}L^{-1}\theta = ML^{-1}T^{-3}$ and $[\rho H] = ML^{-1}T^{-3}$ so that $[H] = L^2T^{-3}$.*

Taking the usual separation of the stress tensor in the thermodynamic stress and the mechanical stress, i.e $\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\pi}$, we get

$$
\rho\frac{De}{Dt} + p\frac{\partial \mathbf{v}_j}{\partial x_j} = \pi_{ij}\frac{\partial \mathbf{v}_i}{\partial x_j} + \frac{\partial}{\partial x_j}(k\frac{\partial T}{\partial x_j}) + \rho H
\tag{18.78}
$$

or

$$
\rho\frac{De}{Dt} + p\vec{\nabla}\cdot\vec{\mathbf{v}} = \boldsymbol{\pi} : \vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\cdot(k\vec{\nabla}T) + \rho H
$$

Note that under the Stokes hypothesis (see before), we have

$$
\rho\frac{De}{Dt} + p\vec{\nabla}\cdot\vec{\mathbf{v}} = \boldsymbol{\tau} : \vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\cdot(k\vec{\nabla}T) + \rho H
$$

From thermodynamic considerations[9] (not treated here) we have the relation $de = TdS - pdv$ between internal energy e, the entropy S, and the specific volume $v = 1/\rho$. This relation is used to transform Eq. (18.78) into the entropy form of the energy equation:

$$
\frac{De}{Dt} = T\frac{DS}{Dt} - p\frac{D(1/\rho)}{Dt} = T\frac{DS}{Dt} + \frac{1}{\rho^2}p\frac{D\rho}{Dt} = T\frac{DS}{Dt} + \frac{1}{\rho^2}p(-\rho\vec{\nabla}\cdot\vec{\mathbf{v}})
\tag{18.79}
$$

where we have used the continuity equation. We then obtain

$$
\rho\frac{De}{Dt} = \rho T\frac{DS}{Dt} - p\vec{\nabla}\cdot\vec{\mathbf{v}}
$$

so that in the end

$$
\rho T\frac{DS}{Dt} = \boldsymbol{\pi} : \vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\cdot(k\vec{\nabla}T) + \rho H
\tag{18.80}
$$

An **adiabatic** state is a state of reversible processes with no heat exchange with surroundings and is defined by S=constant in which case the lhs of (18.80) is 0. This state is not compatible with

---

[9]See for instance https://en.wikipedia.org/wiki/Maxwell_relations

viscous heat dissipation, heat conduction, or internal heat production, as each of these processes would lead to a non-zero contribution at the rhs of (18.80). A fluid particle flowing in an adiabatic mantle assumes at any time the temperature and pressure of the ambient mantle. Therefore there is no conductive heat exchange in the adiabatic state.

Another useful thermodynamic relation is[10]

$$\frac{DS}{Dt} = \left(\frac{\partial S}{\partial T}\right)_p \frac{DT}{Dt} + \left(\frac{\partial S}{\partial p}\right)_T \frac{Dp}{Dt}$$

For the first term we use the definition of the heat capacity $(\mathrm{J\,K^{-1}})$ at constant pressure

$$C_p = T\left(\frac{\partial S}{\partial T}\right)_P$$

while for the second one we use

$$\left(\frac{\partial S}{\partial p}\right)_T = -\left(\frac{\partial (1/\rho)}{\partial T}\right)_p = \frac{1}{\rho^2}\left(\frac{\partial \rho}{\partial T}\right)_p$$

and the definition of the thermal expansion coefficient $(\mathrm{K^{-1}})$

$$\alpha = -\frac{1}{\rho}\left(\frac{\partial \rho}{\partial T}\right)_p$$

so that in the end we have the thermodynamical relationship

$$\begin{aligned}
\frac{DS}{Dt} &= \frac{C_p}{T}\frac{DT}{Dt} - \left(\frac{\partial(1/\rho)}{\partial T}\right)_p \frac{Dp}{Dt} \\
&= \frac{C_p}{T}\frac{DT}{Dt} + \frac{1}{\rho^2}\left(\frac{\partial \rho}{\partial T}\right)_p \frac{Dp}{Dt} \\
&= \frac{C_p}{T}\frac{DT}{Dt} - \frac{\alpha}{\rho}\frac{Dp}{Dt}
\end{aligned}$$

Remark. *Let us look again at the dimensions of these quantities. The entropy S is in $\mathrm{J\,K^{-1}}$, or $[S] = ML^2T^{-2}\theta^{-1}$, and we know that C_p has the same unit. We have $[\alpha] = \theta^{-1}$, $[\rho] = ML^{-3}$ and $[dp] = ML^{-1}T^{-2}$ so we indeed recover $[\alpha dp/\rho] = \theta^{-1}ML^{-1}T^{-2}M^{-1}L^3 =$. All in well.*

We have

$$\frac{DS}{Dt} = \frac{C_p}{T}\frac{DT}{Dt} - \frac{\alpha}{\rho}\frac{Dp}{Dt}$$

Applying this relation to (18.80) leads to the temperature form of the heat equation

$$\rho C_p \frac{DT}{Dt} - \alpha T \frac{Dp}{Dt} = \boldsymbol{\tau} : \vec{\nabla}\vec{v} + \vec{\nabla}\cdot(k\vec{\nabla}T) + \rho H \tag{18.81}$$

The lhs of this equation is zero for an adiabatic state. In this case we find that

$$\frac{dT_a}{dp} = \frac{\alpha T_a}{\rho C_p} \tag{18.82}$$

which gives how temperature changes due to pure adiabatic (de)compression.

In the end we solve the energy equation in this form:

---

[10]https://en.wikipedia.org/wiki/Relations_between_heat_capacities

$$\rho C_p \frac{DT}{Dt} - \vec{\nabla} \cdot k\vec{\nabla}T = \alpha T \frac{Dp}{Dt} + \Phi + \rho H \tag{18.83}$$

where Φ is the shear heating. Note that it couples temperature, velocity (and its derivative the strain rate) and pressure.

In many publications the shear hearing is denoted by Φ and is given by $\Phi = \tau_{ij}\partial_j \mathbf{v}_i = \boldsymbol{\tau} : \vec{\nabla}\vec{\mathbf{v}}$ where $\boldsymbol{\tau}$ is the deviatoric stress tensor. In what follows I use the index notation as it makes for easier derivations:

$$
\begin{aligned}
\Phi &= \tau_{ij}\partial_j \mathbf{v}_i \\
&= 2\eta \dot{\varepsilon}^d_{ij}\partial_j \mathbf{v}_i \\
&= 2\eta \frac{1}{2}\left(\dot{\varepsilon}^d_{ij}\partial_j \mathbf{v}_i + \dot{\varepsilon}^d_{ji}\partial_i u_j\right) \\
&= 2\eta \frac{1}{2}\left(\dot{\varepsilon}^d_{ij}\partial_j \mathbf{v}_i + \dot{\varepsilon}^d_{ij}\partial_i u_j\right) \\
&= 2\eta \dot{\varepsilon}^d_{ij}\frac{1}{2}\left(\partial_j \mathbf{v}_i + \partial_i u_j\right) \\
&= 2\eta \dot{\varepsilon}^d_{ij}\dot{\varepsilon}_{ij} \\
&= 2\eta \dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}} \\
&= 2\eta \dot{\boldsymbol{\varepsilon}}^d : \left(\dot{\boldsymbol{\varepsilon}}^d + \frac{1}{3}(\vec{\nabla}\cdot\vec{\mathbf{v}})\mathbf{1}\right) \\
&= 2\eta \dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d + 2\eta \dot{\boldsymbol{\varepsilon}}^d : \mathbf{1}(\vec{\nabla}\cdot\vec{\mathbf{v}}) \\
&= 2\eta \dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d \tag{18.84}
\end{aligned}
$$

Finally (in Cartesian coordinates)

$$\Phi = \boldsymbol{\tau} : \vec{\nabla}\vec{\mathbf{v}} = 2\eta \dot{\boldsymbol{\varepsilon}}^d : \dot{\boldsymbol{\varepsilon}}^d = 2\eta \left((\dot{\varepsilon}^d_{xx})^2 + (\dot{\varepsilon}^d_{yy})^2 + 2(\dot{\varepsilon}^d_{xy})^2\right) \tag{18.85}$$

See Schubert & Yuen (1978) [1219] for an analysis of shear heating instability in the upper mantle.

Let us quickly look at the $Dp/Dt = \partial_t p + \vec{\mathbf{v}} \cdot \vec{\nabla}p$ term. Often the term $\partial p/\partial t$ is neglected and the pressure is assumed to be mostly hydrostatic in this term so that $\vec{\nabla}p = -\rho\vec{g}$ which yields a much simpler formulation.

See discussions about shear heating ("viscous dissipation") – meaning and application– in Froide-vaux [446] (1973), Stein [1279] (1978), Bird and Yuen [97] (1979), Sleep, Stein, Geller, and Gordon [1251] (1979), Winter [1465] (1987), Masek and Duncan [887] (1998).

A mantle in the adiabatic state satisfies the equilibrium equation $\vec{\nabla}p = \rho\vec{g}$. Assuming spherical symmetry we have $dp/dr = -\rho g$. Substitution in (18.82) gives

$$\frac{dT_a}{dr} = -\frac{\alpha T_a g}{C_p}, \tag{18.86}$$

i.e. the **adiabatic temperature gradient**.

Along similar lines, for a mantle in motion we can approximate the second term on the lhs of (18.81) (the adiabatic (de)compression term) as

$$\alpha T \frac{dp}{dt} \simeq -\alpha T \frac{\rho\vec{g}\cdot d\vec{r}}{dt} = -\alpha\rho T\vec{g}\cdot\vec{\mathbf{v}} \simeq -\alpha\rho T g_r \mathbf{v}_r \tag{18.87}$$

where g_r and $\mathbf{v}_r$ are the radial components of gravity and velocity, respectively.

Exercise: 25. *Determine the temperature as a function of depth for a particle that is being transported through the center of a vertical mantle upwelling. Assume adiabatic conditions and a constant vertical flow velocity.*

Exercise: 26. *Show that for a Newtonian fluid the dissipation term $\Phi = \pi_{ij}\partial v_i/\partial x_j$ can be written as $\Phi = \pi_{ij}\dot{\varepsilon}_{ij} = 2\eta\dot{\varepsilon}'_{ij}\dot{\varepsilon}'_{ij} + \xi(\dot{\varepsilon}_{kk})^2$. What do each term physically account for?*

Exercise: 27. *Demonstrate that for 2D laminar flow (as in exercise 19) the dissipation function can be written as $\Phi = \eta(\partial v_x/\partial z)^2$. Use this to evaluate the dissipative heat production in the flow of exercise 19a assuming zero horizontal pressure gradient (Couette flow). Calculate a numerical value of this heat production using values of $h = 200$ km, $\eta = 10^{21}$ Pa s, $v_0 = 1$ cm yr^{-1}. Compare this to estimates of radiogenic heat production in the upper mantle of $8.4 \cdot 10^{-9}$ J s^{-1} m^{-3}.*

Exercise: 28. *The Navier-Stokes equation for perturbations relative to a hydrostatic reference state is in Cartesian coordinates*

$$\rho_0 \frac{Dv_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} + \tilde{\rho}g_i^0$$

(see equation 18.40). The density perturbation is driving the flow of a medium contained in the material volume V. Assume that gravity is only working vertical $\vec{g} = (0, 0, g)^T$ and that the velocity field is $\vec{v} = (u, v, w)^T$. Assume incompressible flow and that the inertial term can be neglected.

a) Derive the following energy conservation law relating the dissipation of gravitational energy into the energy released due to frictional flow:

$$\int_V \tilde{\rho}gw \, dV = \int_V \pi_{ij}\frac{\partial v_i}{\partial x_j} \, dV$$

Hint: Take the inner product of the equation of motion with the velocity field and integrate the result over V. Assume an impermeable boundary S of V (i.e. $\vec{v} \cdot \vec{n} = 0$) and that the boundary is shear stress free (free slip) or has no-slip ($\vec{v} = \vec{0}$).

b) Show by substitution of the Newtonian rheology that the rhs of this equation is positive and show that the lhs of the equation equals the power due to the change in gravitational potential energy as a result of $\tilde{\rho}$ w.r.t. ρ_0.

Exercise: 29. *Derive the heat equation for a static medium ($\vec{v} = \vec{0}$) from the conservation law of thermal energy. (Hint: Create the heat balance for a fixed control volume based on internal energy, heat production and heat flow through the boundary of the volume).*

18.8   The equation of state

An equation of state relates basic thermodynamic parameters such as density, temperature, pressure, or entropy. For an isochemical fluid, we can choose two independent thermodynamic parameters on which all other depend. Here we choose temperature and pressure as independent parameters, e.g. $\rho(T, p)$. Mantle convection is usually studied relative to some motionless hydrostatic reference state either with a conductive or with an adiabatic geotherm. The perturbations with respect to this state are related to the convective state of the mantle. Therefore, we write $\rho(T, p) = \rho_0(T_0, p_0) + \tilde{\rho}(\tilde{T}, \tilde{p})$ with $T = T_0 + \tilde{T}$ and $p = p_0 + \tilde{p}$ where the 0-subscript denotes reference state quantities and the 'tilde' variables the perturbations with respect to the reference state. The reference state quantities do not depend on time but can depend on the spatial coordinates. The usual assumption is that $\tilde{\rho}(\tilde{T}, \tilde{p})$ depends linearly on the perturbations in temperature and pressure. In this case one can use the thermodynamic relation

$$d\rho = \left(\frac{\partial \rho}{\partial T}\right)_P dT + \left(\frac{\partial \rho}{\partial p}\right)_T dp = -\alpha \rho dT + K_T^{-1} \rho dp$$

where α is the thermal expansion coefficient and K_T the isothermal incompressibility, or isothermal bulk modulus[11] (units: Pa). Assuming linearity, this relation is up-scaled to the macroscopic mantle to obtain the equation of state

$$\rho(\vec{r}, t) = \rho_0(\vec{r})(1 - \alpha\tilde{T} + K_T^{-1}\tilde{p}) \tag{18.88}$$

hence

$$\tilde{\rho} = -\alpha\rho_0\tilde{T} + K_T^{-1}\rho_0\tilde{p}$$

A simple extension to a 2-phase medium consisting of materials with reference densities ρ_0 and ρ_1 is obtained by assuming that the dependence of density of T and p is the same for both phases. Then, only an addition factor is needed to account for the density of a mixed composition

$$\rho(\vec{r}, t) = \rho_0(\vec{r})\left(1 + \Gamma(\frac{\rho_1 - \rho_0}{\rho_0})\right)(1 - \alpha\tilde{T} + K_T^{-1}\tilde{p}) \tag{18.89}$$

Taking $\rho_1 > \rho_0$, the phase distribution Γ can assume values between 0 and 1.

Exercise: 30. *Assume an isothermal (i.e. $\tilde{T} = 0$) and incompressible fluid ($K_T^{-1} = 0$) and derive from the conservation law of mass, the conservation law $D\Gamma/Dt = 0$ for the phase distribution function.*

The reference quantities satisfy equations of a motionless reference state. The definition of the reference state depends on the problem studied. Usually the reference temperature is chosen constant or to follow a mantle adiabat or a conductive geotherm. For a simple reference state with constant thermodynamic parameters the following equations suffice:

$$\vec{\nabla}p_0 = -\rho_0\vec{\nabla}U_0 \tag{18.90}$$
$$\vec{\nabla}^2 U_0 = 4\pi\mathcal{G}\rho_0 \tag{18.91}$$
$$\vec{\nabla}\cdot(k\vec{\nabla}T_0) = 0 \qquad \frac{DT_0}{Dt} = 0 \tag{18.92}$$

More complex reference states involving depth variable thermodynamic parameters require internally consistent relations between thermodynamic parameters such as α, K_T, C_p k and temperature, density, and pressure.

---

[11]The bulk modulus of a substance is a measure of the resistance of a substance to bulk compression. It is defined as the ratio of the infinitesimal pressure increase to the resulting relative decrease of the volume

18.8.1 The complete set of perturbation equations

The following equations (18.93),(18.94),(18.95),(18.96),(18.97) give the relevant equations in terms of the perturbations w.r.t. the adiabatic reference state as defined with equations (18.90), (18.91) and (18.92). Some terms then cancel in the general equations; no approximations have been made.

The equation of state is

$$\rho(\vec{r}, t) = \rho_0(\vec{r}) \left(1 - \alpha \tilde{T} + K_T^{-1} \tilde{p} \right) \tag{18.93}$$

The heat equation in terms of $\tilde{T}$

$$\rho C_p \frac{D\tilde{T}}{Dt} = \frac{\partial}{\partial x_j} \left(k \frac{\partial \tilde{T}}{\partial x_j} \right) + \rho H + \pi_{ij} \frac{\partial \mathbf{v}_i}{\partial x_j} + \alpha T \frac{Dp}{Dt} \tag{18.94}$$

The equation of motion is in terms of $\tilde{p}, \tilde{\rho}, \tilde{T}$:

$$\rho \frac{D\mathbf{v}_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \frac{\partial \pi_{ij}}{\partial x_j} - \rho_0 (\alpha \tilde{T} + K_T^{-1} \tilde{p}) g_i \tag{18.95}$$

The Poisson equation in terms of $\tilde{U}$ and $\tilde{\rho}$:

$$\vec{\nabla}^2 \tilde{U} = 4\pi \mathcal{G} \tilde{\rho} \tag{18.96}$$

and, finally, the continuity equation in terms of $\tilde{\rho}$ reads

$$\frac{\partial \tilde{\rho}}{\partial t} + \frac{\partial (\rho_0 + \tilde{\rho}) \mathbf{v}_j}{\partial x_j} = 0 \tag{18.97}$$

18.9 Scaling of equations

An important subject is the scaling of equations by means of scaling the relevant parameters with appropriate estimates (expected values) for the convection problem at hand. Here, we discuss scaling of parameters for the study of whole mantle convection, however, other studies such as boundary layer modeling or subduction modeling may require different scaling parameters. The idea behind scaling of equations is to find out the relative importance (contribution) of separate terms. Usual scaling parameters for mantle convection are:

$$x_i = x_i' h \qquad t = t' \frac{h^2}{\kappa} \qquad \mathbf{v}_i = \mathbf{v}_i' \frac{\kappa}{h} \qquad \tilde{T} = \tilde{T}' \Delta T_m \qquad \tilde{p} = \tilde{p}' \frac{\eta \kappa}{h^2} \qquad \rho = \rho' \rho_m$$

where h is the mantle thickness, $\kappa = k/(\rho C_p)$ is the thermal diffusivity, ΔT_m is the temperature difference between the Core-Mantle boundary and the surface, η is the dynamic viscosity[12] and ρ_m is a scaling density.

Note that on Earth, we have $\alpha \sim 3{\cdot}10^{-5}$ K^{-1}, $K_T \sim 3{\cdot}10^{12}$ Pa, $\Delta T_m \sim 10^3$ K, $C_p \sim 10^3$ J kg^{-1} K^{-1}, $k \sim 4$ W m^{-1} K^{-1}, $\kappa \sim 1.2 \times 10^{-6}$ m^2 s^{-1}.

The primed quantities are **dimensionless** and are of the order 1 if the scaling is correct for the problem at hand[13]. Application of scaling to the equation of state (18.93) leads to

$$\rho' = \rho_0' \left(1 - \alpha \Delta T_m \tilde{T}' + \frac{K_T^{-1} \eta \kappa}{h^2} \tilde{p}' \right) \tag{18.98}$$

---

[12]Note that the viscosity varies by many orders of magnitude, i.e. between $\sim 10^{17}$ and $\sim 10^{26}$ in the mantle and lithosphere, so finding the appropriate scaling value is necessary.

[13]Except for the viscosity, obviously

The coefficients of the primed-quantities determine the relative importance of the terms (with respect to 1). We have $\alpha\Delta T_m \sim 3\cdot 10^{-2}$ and $K_T^{-1}\eta\kappa/h^2 \sim 2.5\cdot 10^{-10}$, which leads us to conclude that thermal perturbations of density are much more important than pressure perturbations (assuming the equation of state and scaling are correct for the convection problem). We note that the scaling factor of pressure is about 100 Pa, i.e. perturbations of hydrostatic pressure are expected to be small compared to the rheological stress (1-100 MPa).

The coefficient $K_T^{-1}\eta\kappa/h^2$ can be written as $\eta C_p/k \cdot k^2 K_T^{-1}/C_p^2/\rho_m h^2 = \mathsf{Pr}\cdot\mathsf{M}^2$ where where Pr is the Prandtl number[14] and M the Mach number[15]. M gives the ratio between flow speed and sound speed which is about 10^{-16} for mantle convection. The Prandtl number is about 10^{24} and will appear in the coefficient of the inertial term of the dimensionless equation of motion.

Let us now turn to the scaling of the continuity equation (18.97) which leads to some basic insight into the problem of (in-)compressibility. The dimensionless version is

$$\frac{\partial\tilde{\rho}'}{\partial t'} + \frac{\partial(\rho_0'+\tilde{\rho})\mathsf{v}_j'}{\partial x_j'} = 0 \qquad ; \qquad \rho_0' = \frac{\rho_0(\vec{r})}{\rho_m} \tag{18.99}$$

Exercise: 32. *Derive Eq.* (18.99).

Substitution of the equation of state (18.93) and taking the limit $Pr\cdot M^2\to\infty$ leads to

$$-\rho_0'\alpha\Delta T_m\frac{\partial\tilde{T}'}{\partial t'} + \frac{\partial}{\partial x_j'}\left[(\rho_0' - \rho_0'\alpha\Delta T_m\tilde{T}')\mathsf{v}_j'\right] = 0 \tag{18.100}$$

which gives the conservation law based on thermal perturbations of density only. In the limit that the perturbation terms are very small we arrive at

$$\frac{\partial}{\partial x_j'}(\rho_0'\mathsf{v}_j') = 0 \tag{18.101}$$

which is the equation of anelastic conservation of mass. Compared to (18.97) we effectively replaced the density by the reference density. The time derivative of density has disappeared. This derivative is primarily related to the propagation of seismic waves. This process occurs at totally different time scales than mantle convection. The anelastic conservation of mass (18.101) is being used in convection modeling of compressible fluids, see for instance the equations of the ASPECT code[16]:

Specifically, we consider the following set of equations for velocity $\mathbf{u}$, pressure p and temperature T, as well as a set of advected quantities c_i that we call *compositional fields*:

$$-\nabla\cdot\left[2\eta\left(\varepsilon(\mathbf{u}) - \frac{1}{3}(\nabla\cdot\mathbf{u})\mathbf{1}\right)\right] + \nabla p = \rho\mathbf{g} \text{ in }\Omega, \tag{1}$$

$$\nabla\cdot(\rho\mathbf{u}) = 0 \text{ in }\Omega, \tag{2}$$

$$\rho C_p\left(\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T\right) - \nabla\cdot k\nabla T = \rho H$$
$$+ 2\eta\left(\varepsilon(\mathbf{u}) - \frac{1}{3}(\nabla\cdot\mathbf{u})\mathbf{1}\right):\left(\varepsilon(\mathbf{u}) - \frac{1}{3}(\nabla\cdot\mathbf{u})\mathbf{1}\right)$$
$$+ \alpha T\left(\mathbf{u}\cdot\nabla p\right)$$
$$+ \rho T\Delta S\left(\frac{\partial X}{\partial t} + \mathbf{u}\cdot\nabla X\right) \text{ in }\Omega, \tag{3}$$

$$\frac{\partial c_i}{\partial t} + \mathbf{u}\cdot\nabla c_i = q_i \text{ in }\Omega, i = 1\ldots C \tag{4}$$

where $\varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$ is the symmetric gradient of the velocity (often called the *strain rate*)[1].

In this set of equations, (1) and (2) represent the compressible Stokes equations in which $\mathbf{u} = \mathbf{u}(\mathbf{x},t)$ is the velocity field and $p = p(\mathbf{x},t)$ the pressure field. Both fields depend on space $\mathbf{x}$ and time t. Fluid flow is driven by the gravity force that acts on the fluid and that is proportional to both the density of the fluid and the strength of the gravitational pull.

Taken from the ASPECT website. Note that the equations are not dimensionless.

---

[14]https://en.wikipedia.org/wiki/Prandtl_number

[15]https://en.wikipedia.org/wiki/Mach_number

[16]https://aspect-documentation.readthedocs.io/en/latest/user/methods/basic-equations/index.html

In Solheim and Peltier [1256] we read:

"The anelastic-liquid approximation [e.g., Jarvis and McKenzie [672] (1980)] is applied to the system [mass, momentum, energy conservation equations]. This involves setting $\partial \rho / \partial t = 0$ in the mass conservation equation, assuming g, C_p, α, k, K_T and κ to be known functions of radius and replacing ρ by ρ_r everywhere in [the equations] except in the body force term of the momentum equation. Because the mantle has essentially infinite Prandtl number, the inertial force term in the momentum conservation equation may be neglected [1047]. Furthermore, owing to the extremely small velocities associated with the mantle convection process, the pressure distribution is very nearly that of a fluid in hydrostatic equilibrium and we may then safely assume

$$\alpha T \frac{Dp}{Dt} \simeq -\alpha T \rho g u_r$$

where u_r is the radial component of the velocity. These approximations have been discussedin greater detail by Solheim and Peltier [1258] (1990). "

18.10   The Boussinesq approximation

The Boussinesq[17] approximation is ubiquitous in computational geodynamics [1266, 1488]. This approximation leads to a set of simplified equations that are easier to solve for some analytical problems and in numerical modeling of convective flow. The following approximations are being made:

- Neglect the effect of density variations with respect to a reference state except in terms related to the driving force of convection

- assume that the divergence of the velocity field is 0 (in this case $\boldsymbol{\pi} = \boldsymbol{\tau}$).

- Only concern temperature changes resulting from diffusion and advection (i.e. neglect terms related to adiabatic compression, and heat dissipation)

This simplifies Eqs. (18.93),(18.94),(18.95),(18.96),(18.97) to

$$\rho(\vec{r}, t) = \rho_0 (1 - \alpha \tilde{T}) \tag{18.102}$$

$$\rho_0 C_p \frac{D\tilde{T}}{Dt} = \frac{\partial}{\partial x_j} \left(k \frac{\partial \tilde{T}}{\partial x_j} \right) + \rho_0 H \tag{18.103}$$

$$\rho_0 \frac{D\boldsymbol{\nu}_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} - \rho_0 \alpha \tilde{T} g_i \tag{18.104}$$

$$\vec{\nabla}^2 \tilde{U} = -4\pi \mathcal{G} \rho_0 \alpha \tilde{T} \tag{18.105}$$

$$\frac{\partial \boldsymbol{\nu}_j}{\partial x_j} = 0 \tag{18.106}$$

18.10.1   The Rayleigh-Bénard convection

This is the classical example of a laterally unlimited fluid layer of thickness h in a gravity field. We derive here the pertinent equations in the Boussinesq approximation. In the next section these equations are being used to study the problem of onset of convection.

---

[17]https://en.wikipedia.org/wiki/Joseph_Valentin_Boussinesq
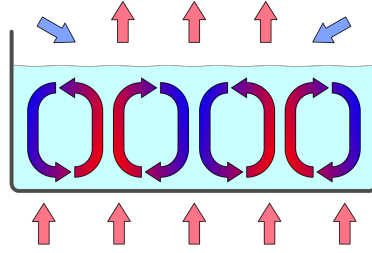
Convection cells in a gravity field. Taken from `https://en.wikipedia.org/wiki/Rayleigh_Benard_convection`

Gravity is directed along the positive z-axis. The top and bottom of the layer are kept at a constant temperature:

$$T(x, z = 0, t) = T_0 \quad \text{and} \quad T(x, z = h, t) = T_0 + \Delta T \quad (\Delta T > 0) \tag{18.107}$$

The mechanical boundary conditions on the top and bottom are impermeability and shear stress free (also called free slip):

$$
\begin{aligned}
\boldsymbol{\nu}_z(x, z = 0, t) &= 0 & (18.108) \\
\boldsymbol{\nu}_z(x, z = h, t) &= 0 & (18.109) \\
\tau_{xz}(x, z = 0, t) &= 0 & (18.110) \\
\tau_{xz}(x, z = h, t) &= 0 & (18.111) \\
\tau_{yz}(x, z = 0, t) &= 0 & (18.112) \\
\tau_{yz}(x, z = h, t) &= 0 & (18.113)
\end{aligned}
$$

Exercise: 33. *Show for an incompressible Newtonian fluid that at $z = 0$ the traction $\vec{t}^h$ is given by $t_x^n = t_y^n = 0$ and $t_z^n = -p + 2\eta \partial \boldsymbol{\nu}_z / \partial z$.*

The condition of zero vertical velocity is only approximately valid at the Earth's surface. Vertical motions can now be measured by geodetic techniques and are generally one order of magnitude less than horizontal motions. Also from the geological past we know that horizontal motions have had much larger amplitude than vertical motions. If the Earth would have a pure fluid behavior (has it?) then mantle flow can induce **dynamic surface topography**. The vertical velocity $\boldsymbol{\nu}_z$ equals the local time derivative $\partial d / \partial t$ of the vertical surface deflection $d(t)$. In numerical solutions (with impermeable boundaries along which pressure variation can accumulate) the deflection is often computed a posteriori by equating $\rho g d \sim -p + 2\eta \partial \boldsymbol{\nu}_z / \partial z$.
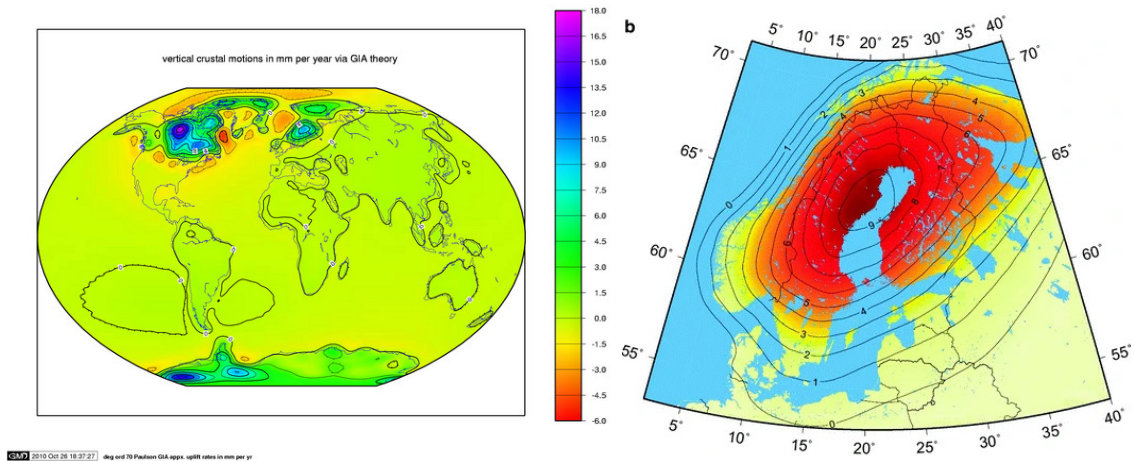
In the following we assume an isoviscous Newtonian fluid with constant thermal conductivity, constant heat capacity and no heat production and we adopt the Boussinesq approximation. Then we have the following equations for Rayleigh-Benard convection:

$$\tilde{\rho} = -\rho_0 \alpha \tilde{T} \tag{18.114}$$

$$\frac{D\tilde{T}}{Dt} = \kappa \vec{\nabla}^2 \tilde{T} \tag{18.115}$$

$$\rho_0 \frac{D\mathbf{v}_i}{Dt} = -\frac{\partial \tilde{p}}{\partial x_i} + \eta \vec{\nabla}^2 \mathbf{v}_i - \rho_0 \alpha \tilde{T} g_i \tag{18.116}$$

$$\vec{\nabla}^2 \tilde{U} = -4\pi \mathcal{G} \rho_0 \alpha \tilde{T} \tag{18.117}$$

$$\frac{\partial \mathbf{v}_j}{\partial x_j} = 0 \tag{18.118}$$

The reference temperature is assumed to be constant or following an adiabat or is a time stationary conductive geotherm. Reference pressure is computed from the hydrostatic equation $\vec{\nabla}p = \rho_0 \vec{g}$.

We assume a 2-D situation in which the velocity vector is denoted by $\vec{\mathbf{v}} = (u, 0, w)^T$. The mechanical boundary condition becomes $\tau_{xz}(x, z = 0) = \tau_{xz}(x, z = h) = 0$ and $\eta \partial u / \partial z(x, z = 0) = \eta \partial u / \partial z(x, z = h) = 0$. The scaling of the equation of motion gives

$$\frac{1}{\mathsf{Pr}} \frac{D\mathbf{v}_i'}{Dt'} = -\frac{\partial \tilde{p}'}{\partial x_i'} + (\nabla')^2 \mathbf{v}_i' - \mathsf{Ra}\tilde{T}' \delta_{zi} \tag{18.119}$$

where the Prandtl number is $\mathsf{Pr} = \eta C_p / k = \eta / \rho_0 \kappa$ and the Rayleigh number is $\mathsf{Ra} = \rho_0 \alpha g \Delta T h^3 / \eta \kappa$. For the Earth's mantle $Pr \sim 10^{24}$ and $\mathsf{Ra} \sim 10^7$ which leads to the conclusion that the inertial term can be neglected. Please check Section 3.11.2 for the complete adimensionalisation of the temperature-dependent Navier-Stokes equations. We arrive at the non-dimensional equation (dropping the primes for convenience):

$$0 = -\vec{\nabla}\tilde{p} + \vec{\nabla}^2 \vec{\mathbf{v}} - \mathsf{Ra}\tilde{T}\vec{e}_z \tag{18.120}$$

in which Ra is the only free parameter. Recall that $\tilde{p}$ is the pressure anomaly with regards to p_0. The Rayleigh number determines the magnitude of the force driving convection (see computer practical).

Because we considerer a 2-D situation of an incompressible fluid we can adopt the stream function approach with $u = \partial \Psi / \partial z$ and $w = -\partial \Psi / \partial x$ which leads to

$$\vec{\nabla}^4 \Psi = -\mathsf{Ra}\frac{\partial \tilde{T}}{\partial x} \tag{18.121}$$

(to be compared with (18.48)) with boundary conditions $\Psi(x, z = 0) = \Psi(x, 1) = 0$ and $\partial^2 \Psi / \partial z^2(x, y = 0) = \partial^2 \Psi / \partial z^2(x, y = 1) = 0$.

The energy equation becomes after scaling

$$\frac{D\tilde{T}}{Dt} = \vec{\nabla}^2 \tilde{T} \tag{18.122}$$

---

[18] https://en.wikipedia.org/wiki/Post-glacial_rebound

Equation (18.121) and (18.122) are coupled non-linear equations which usually need to be solved numerically.

Recall that the temperature $\tilde{T}$ is the deviation from either a constant temperature of from adiabatic temperature $\tilde{T} = T - T_a$. Advective transport of adiabatic temperature does not lead to temperature changes with the surroundings and therefore $\tilde{T}$ is the temperature associated with convective flow. In convective flow that assumes a (non-adiabatic) conductive geotherm in the background we can separate $\tilde{T}$ as $\tilde{T} = \tilde{T}_0 + \tilde{T}_1$ where $\tilde{T}_0$, is the geotherm of the conductive reference state and $\tilde{T}_1$ is the perturbation of the temperature field.

Exercise: 34. *Derive the stationary conductive reference temperature profile $\tilde{T}_0'(z') = z'$ using the temperature boundary conditions given earlier (with $T_{surface} = 0$).*

18.10.2 Linear stability analysis (the onset of convection problem)

We are now ready to solve a linear stability problem which will provide fundamental insight in the role of the Rayleigh number in convection.

The system is a layer of fluid between $y = 0$ and $y = h$, with boundary conditions $T(x, y = 0) = T_b$ and $T(x, y = h) = 0$, characterized by ρ_0, C_p, k, η_0 which are assumed to be constant (in space and time).

The Stokes equation is $\vec{\nabla} \cdot \boldsymbol{\sigma} + \rho \vec{g} = \vec{0}$. The components of the this equation on the x- and $y-$axis are:

$$
\begin{aligned}
(\vec{\nabla} \cdot \boldsymbol{\sigma})_x &= -\rho \vec{g} \cdot \vec{e}_x = 0 \\
(\vec{\nabla} \cdot \boldsymbol{\sigma})_y &= -\rho \vec{g} \cdot \vec{e}_y = \rho g_0
\end{aligned}
$$

since $\vec{g}$ and $\vec{e}_y$ are in opposite directions ($\vec{g} = -g_0 \vec{e}_y$, with $g_0 > 0$).

Following Eq. (18.48), the stream function formulation of the incompressible isoviscous Stokes equation is

$$
\eta_0 \nabla^4 \Psi = \frac{\partial \rho g_y}{\partial x} - \frac{\partial \rho g_x}{\partial y} = \frac{\partial \rho g_y}{\partial x} = -g_0 \frac{\partial \rho}{\partial x}
$$

since $g_x = 0$ and $g_y = \vec{g} \cdot \vec{e}_y = -g_0$. Assuming a linearised density field with regards to temperature $\rho(T) = \rho_0(1 - \alpha T)$ we have

$$
\frac{\partial \rho}{\partial x} = -\rho_0 \alpha \frac{\partial T}{\partial x}
$$

and then

$$
\vec{\nabla}^4 \Psi = \frac{\rho_0 g_0 \alpha}{\eta_0} \frac{\partial T}{\partial x} \tag{18.123}
$$

For small perturbations of the conductive state[19] $T_c(y) = (1 - y/h)T_b$ we define the temperature perturbation $\tilde{T}(x, y)$ such that

$$
T(x, y, t) = T_c(y) + \tilde{T}(x, y, t)
$$

Note that the temperature perturbation $\tilde{T}$ must satisfy the homogeneous boundary conditions $\tilde{T}(x, y = 0) = 0$ and $\tilde{T}(x, y = h) = 0$. We then have[20]:

$$
\boxed{\vec{\nabla}^4 \Psi = \frac{\rho_0 g_0 \alpha}{\eta_0} \frac{\partial \tilde{T}}{\partial x}} \tag{18.124}
$$

---

[19]The conductive state temperature is defined as the solution of the steady state diffusion equation $\Delta T_c = 0$ subjected to the desired boundary conditions at the top and at the bottom.

[20]This is the same equation as in Turcotte & Schubert, eq 6.310.

In the absence of heat production, the temperature equation (in the Boussinesq approx.) is

$$\rho_0 C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = k \Delta T \quad \Rightarrow \quad \rho_0 C_p \left(\frac{\partial (T_c + \tilde{T})}{\partial t} + \vec{v} \cdot \vec{\nabla}(T_c + \tilde{T}) \right) = k \Delta (T_c + \tilde{T}) \tag{18.125}$$

First we start by acknowledging that T_c does not depend on time, so that $\partial T_c / \partial t = 0$. Then, $\Delta T_c = 0$ since T_c is a linear function of y. Finally, we assume the nonlinear term $\vec{v} \cdot \vec{\nabla} \tilde{T}$ to be second order (temperature perturbations and coupled velocity changes are assumed to be small). In the end, defining the heat diffusion coefficient κ as $\kappa = k/\rho_0 C_p$, the energy equation can be simplified further as follows:

$$\frac{\partial \tilde{T}}{\partial t} + \vec{v} \cdot \vec{\nabla} T_c = \kappa \Delta \tilde{T}$$

Using the relationship between velocity and stream function $\vec{v} = (u, v) = (\partial_y \Psi, -\partial_x \Psi)$ and since $\vec{\nabla} T_c = -(T_b/h)\vec{e}_y$ then

$$\vec{v} \cdot \vec{\nabla} T_c = \begin{pmatrix} \partial_y \Psi \\ -\partial_x \Psi \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -T_b/h \end{pmatrix} = \frac{T_b}{h} \frac{\partial \Psi}{\partial x}$$

and finally[21]:

$$\boxed{\frac{\partial \tilde{T}}{\partial t} - \kappa \Delta \tilde{T} = -\frac{T_b}{h} \frac{\partial \Psi}{\partial x}} \tag{18.126}$$

Looking at these equations, we immediately think about a separation of variables approach to solve these equations. Both equations showcase the Laplace operator Δ, and the eigenfunctions of the biharmonic operator and the Laplace operator are the same. We then pose that Ψ and $\tilde{T}$ can be written[22]:

$$\tilde{T}(x, y, t) = \tilde{T}_0 \exp(pt) \left[a_k \cos(k_x x) + b_k \sin(k_x x) \right] \left[c_k \cos(k_y y) + d_k \sin(k_y y) \right]$$

$$\Psi(x, y, t) = \Psi_0 \exp(pt) \left[\alpha_k \cos(k_x x) + \beta_k \sin(k_x x) \right] \left[\delta_k \cos(k_y y) + \gamma_k \sin(k_y y) \right]$$

where $\tilde{T}_0$ and Ψ_0, a_k, b_k, c_k, d_k and $\alpha_k, \beta_b, \delta_k, \gamma_k$ are constants. We then of course have

$$\begin{aligned} \vec{\nabla}^2 \tilde{T} &= \frac{\partial^2 \tilde{T}}{\partial x^2} + \frac{\partial^2 \tilde{T}}{\partial y^2} \\ &= \tilde{T}_0 \exp(pt) \left\{ \left[-k_x^2 a_k \cos(k_x x) - k_x^2 b_k \sin(k_x x) \right] \left[c_k \cos(k_y y) + d_k \sin(k_y y) \right] \right\} \\ &+ \tilde{T}_0 \exp(pt) \left\{ \left[a_k \cos(k_x x) + b_k \sin(k_x x) \right] \left[-k_y^2 c_k \cos(k_y y) - k_y^2 d_k \sin(k_y y) \right] \right\} \\ &= -(k_x^2 + k_y^2)\tilde{T} \end{aligned} \tag{18.127}$$

and a similar expression for Ψ.

The boundary conditions on $\tilde{T}$ are $\tilde{T}(y = 0) = \tilde{T}(y = h) = 0$. From the first one it follows immediately that $c_k = 0$. From the second, we arrive at $k_y h = n\pi$, which yields $\sin(n\pi y/h)$ where n is an integer. We then arrive at the following expression for the temperature $\tilde{T}$:

$$\tilde{T}(x, y, t) = \tilde{T}_0 \exp(pt) \left[a_k \cos(k_x x) + b_k \sin(k_x x) \right] \sin\left(n\pi \frac{y}{h} \right)$$

where n is an integer number.

At this stage we make an important assumption: at $t = 0$ we then only consider a single horizontal periodic perturbance with wavelength λ as depicted below. In such a case we expect that it would in fact lead to the formation of the following convection cells:

---

[21]This is the same equation as in Turcotte & Schubert, eq 6.309.

[22]T & S actually very much postulate the final form of these quantities without (enough?) justification. I would like to revisit this in the future and better support this assertion.
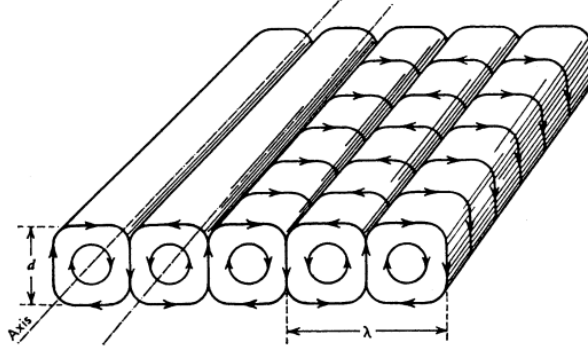
Note that the number of cells left and right of those shown is infinite. Source unknown. The coordinate $x = 0$ is set to the left vertical dashed line for convenience.

The boundary conditions are free slip at the top and at the bottom, i.e. $v(y = 0) = v(y = h) = 0$. Also, by symmetry of the perturbance we see that $u = 0$ at each 'side' (i.e. on the vertical dashed lines of the figure above), i.e. for $x = 0$ and $x = \lambda$.

Let us now turn to the vertical y component of the velocity:

$$
\begin{aligned}
v &= -\frac{\partial \Psi}{\partial x} \\
&= -\frac{\partial}{\partial x} \left\{ \Psi_0 \exp(pt) \left[\alpha_k \cos(k_x x) + \beta_k \sin(k_x x) \right] \left[\delta_k \cos(k_y y) + \gamma_k \sin(k_y y) \right] \right\} \\
&= -\Psi_0 \exp(pt) \left[-\alpha_k k_x \sin(k_x x) + \beta_k k_x \cos(k_x x) \right] \left[\delta_k \cos(k_y y) + \gamma_k \sin(k_y y) \right] \quad (18.128)
\end{aligned}
$$

The boundary condition at the bottom is $v(y = 0) = 0$, so that $\delta_k = 0$ here again. The boundary condition at the top is $v(y = h) = 0$, so that $k_y h = n\pi = 0$ as before. Then

$$
\Psi(x, y, t) = \Psi_0 \exp(pt) \left[\alpha_k \cos(k_x x) + \beta_k \sin(k_x x) \right] \sin\left(n\pi \frac{y}{h} \right)
$$

Turning now to the horizontal component of the velocity:

$$
\begin{aligned}
u &= \frac{\partial \Psi}{\partial y} \\
&= \frac{\partial}{\partial y} \left\{ \Psi_0 \exp(pt) \left[\alpha_k \cos(k_x x) + \beta_k \sin(k_x x) \right] \sin\left(n\pi \frac{y}{h} \right) \right\} \\
&= \Psi_0 \exp(pt) \left[\alpha_k \cos(k_x x) + \beta_k \sin(k_x x) \right] \frac{n\pi}{h} \sin\left(n\pi \frac{y}{h} \right) \quad (18.129)
\end{aligned}
$$

Using now the 'side' boundary conditions: $u(x = 0) = 0$ yields $\alpha_k = 0$ and $u(x = \lambda) = 0$ yields $k_y \lambda = 2\pi$ so that in the end:

$$
\boxed{ \Psi(x, y, t) = \Psi_0 \exp(pt) \sin\left(\frac{2\pi}{\lambda} x \right) \sin\left(\frac{n\pi}{h} y \right) } \quad (18.130)
$$

Looking at the biharmonic equation (18.126), its rhs is $\sim \frac{\partial \Psi}{\partial x}$. Then, the x dependency of this term will be $\cos(2\pi x/\lambda)$. The lhs term of (18.126) is proportional to $\tilde{T}$ (see Eq. (18.127)), i.e. proportional to $a_k \cos(k_x x) + b_k \sin(k_x x)$. For these equations to be compatible, we must set $b_k = 0$ and we then obtain[23]

---

[23]Taking $n = 1$ and remembering that Turcotte & Schubert have the domain between $y - h/2$ and $y = h/2$, these expressions are identical to Eqs. 6.311 and 6.312 of the book. Also one could have assigned $\partial T/\partial x$ on the sides for symmetry reasons and have obtained the same expression.

940

$$\tilde{T}(x, y, t) = \tilde{T}_0 \exp(pt) \cos\left(\frac{2\pi}{\lambda} x\right) \sin\left(n\pi \frac{y}{h}\right) \qquad (18.131)$$

where a_k has been 'absorbed' in $\tilde{T}_0$.

Then the two framed PDEs above, Eq. (18.124) and Eq. (18.126), when coupled with Eq. (18.130) and Eq. (18.131), become:

$$\nabla^4 \Psi = \frac{\rho_0 g_0 \alpha}{\eta_0} \frac{\partial \tilde{T}}{\partial x}$$

$$\Rightarrow \quad \nabla^2 \left[\left(-\frac{4\pi^2}{\lambda^2} - \frac{n^2\pi^2}{h^2}\right)\Psi\right] = \frac{\rho_0 g_0 \alpha}{\eta_0} \frac{\partial \tilde{T}}{\partial x}$$

$$\Rightarrow \quad \left(-\frac{4\pi^2}{\lambda^2} - \frac{n^2\pi^2}{h^2}\right)^2 \Psi = \frac{\rho_0 g_0 \alpha}{\eta_0} \frac{\partial \tilde{T}}{\partial x}$$

$$\Rightarrow \quad \left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 \Psi_0 \exp(pt) \sin\left(\frac{2\pi}{\lambda}x\right) \sin\left(n\pi\frac{y}{h}\right) = \frac{\rho_0 g_0 \alpha}{\eta_0} \cdot -\frac{2\pi}{\lambda}\tilde{T}_0 \exp(pt) \sin\left(\frac{2\pi}{\lambda}x\right) \sin\left(n\pi\frac{y}{h}\right)$$

$$\Rightarrow \quad \left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 \Psi_0 = -\frac{\rho_0 g_0 \alpha}{\eta_0}\frac{2\pi}{\lambda}\tilde{T}_0 \qquad (18.132)$$

$$\frac{\partial \tilde{T}}{\partial t} - \kappa \Delta \tilde{T} = -\frac{T_b}{h}\frac{\partial \Psi}{\partial x}$$

$$\Rightarrow \quad p\tilde{T} - \kappa\left(-\frac{4\pi^2}{\lambda^2} - \frac{n^2\pi^2}{h^2}\right)\tilde{T} = -\frac{T_b}{h}\cdot\frac{2\pi}{\lambda}\Psi_0 \exp(pt)\cos\left(\frac{2\pi}{\lambda}x\right)\sin\left(n\pi\frac{y}{h}\right)$$

$$\Rightarrow \quad \left[p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)\right]\tilde{T}_0 \exp(pt)\cos\left(\frac{2\pi}{\lambda}x\right)\sin\left(n\pi\frac{y}{h}\right) = -\frac{T_b}{h}\frac{2\pi}{\lambda}\Psi_0 \exp(pt)\cos\left(\frac{2\pi}{\lambda}x\right)\sin\left(n\pi\frac{y}{h}\right)$$

$$\Rightarrow \quad \left[p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)\right]\tilde{T}_0 = -\frac{T_b}{h}\frac{2\pi}{\lambda}\Psi_0 \qquad (18.133)$$

We are then left with two equations:

$$\left[p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)\right]\tilde{T}_0 = -\frac{T_b}{h}\frac{2\pi}{\lambda}\Psi_0$$

$$\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 \Psi_0 = -\frac{\rho_0 g_0 \alpha}{\eta_0}\frac{2\pi}{\lambda}\tilde{T}_0$$

which we can cast as

$$\begin{pmatrix} p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right) & \frac{T_b}{h}\frac{2\pi}{\lambda} \\ -\frac{\rho_0 g_0 \alpha}{\eta_0}\frac{2\pi}{\lambda} & -\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 \end{pmatrix}\begin{pmatrix} \tilde{T}_0 \\ \Psi_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The determinant of the matrix should be zero to have non-trivial solutions[24] for the amplitude factors

---

[24]Let us consider the following matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\cdot\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Rightarrow \qquad \begin{pmatrix} ac & bc \\ 0 & ad - bc \end{pmatrix}\cdot\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where we have multiplied the first row by c and the second row by a and subtract row 1 from row 2. The lower right term $ad - bc$ is the determinant and we find that it must be equal to zero since y is not zero.

(i.e. $\tilde{T}_0 = 0$ and $\Psi_0 = 0$ which is not helpful). This leads to the condition:

$$
\begin{aligned}
Det \;&=\; \left[p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)\right] \cdot -\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 + \frac{\rho_0 g_0 \alpha}{\eta_0}\frac{2\pi}{\lambda} \cdot \frac{T_b}{h}\frac{2\pi}{\lambda} \\
&=\; -\left[p + \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)\right]\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 + \frac{\rho_0 g_0 \alpha T_b}{h\eta_0}\frac{4\pi^2}{\lambda^2} \\
&=\; -p\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 - \kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \frac{\rho_0 g_0 \alpha T_b}{h\eta_0}\frac{4\pi^2}{\lambda^2}
\end{aligned}
$$

The determinant is zero for

$$
\begin{aligned}
p\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2 \;&=\; -\kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \frac{\rho_0 g_0 \alpha T_b}{h\eta_0}\frac{4\pi^2}{\lambda^2} \\
p \;&=\; \frac{-\kappa\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \frac{\rho_0 g_0 \alpha T_b}{h\eta_0}\frac{4\pi^2}{\lambda^2}}{\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2} \\
&=\; \kappa\frac{-\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \frac{\rho_0 g_0 \alpha T_b}{h\kappa\eta_0}\frac{4\pi^2}{\lambda^2}}{\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2} \\
&=\; \frac{\kappa}{h^6}\frac{-h^6\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \frac{\rho_0 g_0 \alpha T_b h^3}{\kappa\eta_0}\frac{4\pi^2 h^2}{\lambda^2}}{\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2} \\
&=\; \frac{\kappa}{h^2}\frac{-h^6\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^3 + \mathsf{Ra}\frac{4\pi^2 h^2}{\lambda^2}}{h^4\left(\frac{4\pi^2}{\lambda^2} + \frac{n^2\pi^2}{h^2}\right)^2} \\
&=\; \frac{\kappa}{h^2}\frac{-\left(\frac{4\pi^2 h^2}{\lambda^2} + n^2\pi^2\right)^3 + \mathsf{Ra}\frac{4\pi^2 h^2}{\lambda^2}}{\left(\frac{4\pi^2 h^2}{\lambda^2} + n^2\pi^2\right)^2}
\end{aligned}
\tag{18.134}
$$

where we have used the Rayleigh number of the system defined as

$$
\mathsf{Ra} = \frac{\rho_0 g_0 \alpha T_b h^3}{\eta_0 \kappa}
$$

The coefficient p inside $\exp(pt)$ present in both temperature and stream function expressions determines the stability of the system: if it is negative, the system is stable and both Ψ and $\tilde{T}$ will decay to zero (return to conductive state). If $p = 0$, then the system is meta-stable, and if $p > 0$ then the system is unstable and the perturbations will grow.

In case of the stable regime an initial temperature perturbation will die out (e.g. because conduction wins from advection). In the case of the unstable regime convection will occur with an exponential growth factor. The intermediate, marginally stable, regime is the transition between convection and no convection for which our linearization applies.

The threshold is then $p = 0$ and the corresponding critical Rayleigh number Ra_c is[25]:

$$
\mathsf{Ra}_c = \frac{\left(\frac{4\pi^2 h^2}{\lambda^2} + n^2\pi^2\right)^3}{\frac{4\pi^2 h^2}{\lambda^2}}
$$

---

[25] this is eq 6.319 of T&S for n=1

942

Let us denote $\underline{h} = 2\pi h / \lambda$ the dimensionless thickness of the layer. The critical Rayleigh number is then a function of $\underline{h}$:

$$\mathsf{Ra}_c(\underline{h}) = \frac{(\underline{h}^2 + n^2\pi^2)^3}{\underline{h}^2}$$

It is plotted on the following figure for $n = 1$:
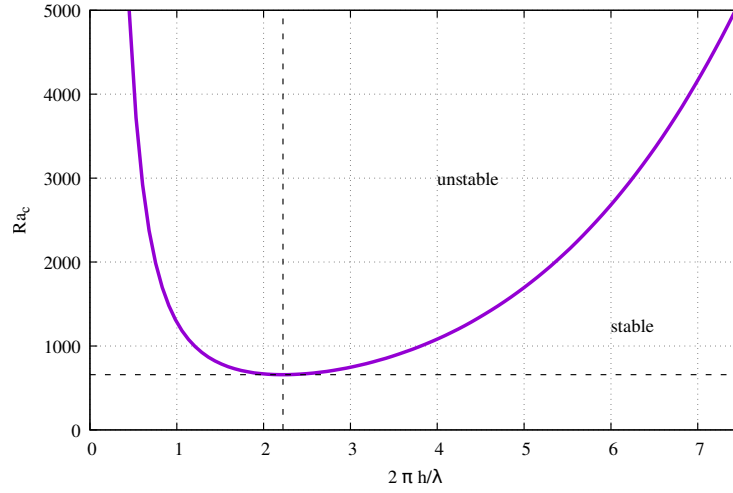
Critical Rayleigh number Ra_c for the onset of convection in a layer heated from below with stress-free boundaries as a function of dimensionless wavenumber $2\pi h/\lambda$ and for $n = 1$. For a system with $\mathsf{Ra} = 2000$ then convection cannot occur for $2\pi h/\lambda < 0.8$ and $2\pi h/\lambda > 5.4$. The dashed lines indicate the minimum critical Rayleigh number and its corresponding $\underline{h}$ value. Unstable means that perturbations will grow and yield convection, while stable means that perturbations will diffuse away. Gnuplot script in `images/chapter_md`.

The minimum critical Rayleigh number is given by

$$\left.\frac{\partial \mathsf{Ra}_c}{\partial (2\pi h/\lambda)}\right|_{n=1} = 0$$

We find[26] that the value of the wavelength corresponding to the smallest value of the critical Rayleigh number is $\lambda = 2\sqrt{2}h$, or $\underline{h} = \pi/\sqrt{2} \simeq 2.22$ and substitution of this value for the wavelength gives the critical Rayleigh number

$$\mathsf{Ra}_c = \frac{27}{4}\pi^4 \simeq 657.5$$

This solves the linearised onset of convection problem in the sense that an unstable layering (cold above hot) only starts convecting after a critical Rayleigh number has been overcome, e.g., by an increased ΔT

These numbers hold for a model with boundaries that are isothermal, impermeable, and free slip. Adopting other boundary conditions leads to different critical numbers. For instance, in the extreme of having fixed (no slip) boundaries one obtains $\mathsf{Ra} \simeq 1707.8$ and $\lambda \simeq 2.016h$ demonstrating that it is more difficult to initiate convection compared to free slip boundaries.

When conducting a similar analysis in a spherical shell, minimum critical Rayleigh numbers prove to be much larger. Free slip (rigid) conditions at the surface and bottom 'CMB' boundary lead to $\mathsf{Ra}_{c,min} \sim 14,000(35,000)$ with a critical wave length of spherical harmonic degree $L = 3(4)$. As the main difference between a flat layer and a spherical shell is the geometry, apparently, convection in a spherical shell experiences strong geometrical constraints (less "space" to flow near the bottom than near the top of the layer and more cooling at the surface compared to less heat input at the bottom).

For realistic values of the physical parameters defining the Rayleigh number and a realistic layer thickness, the only quantity that changes the Rayleigh number is the temperature difference ΔT

---

[26]eq 6.320 of T&S

between top and bottom. **The critical minimum Rayleigh number thus determines the critical ΔT below which no convection occurs and above which convection is enhanced**. The analysis above is only valid in the linear regime, i.e. near the critical Rayleigh number.

Estimates for the Rayleigh number of the Earth's mantle vary between $5 \cdot 10^5$ (upper mantle) to $6 \cdot 10^7$ for the whole mantle which is by many factors larger than the minimum critical Rayleigh numbers that follow from experiments as described above. The mantle is in a state of vigorous convection (on the geological time scale). Thermal expansion and thermal diffusivity are decreasing with depth while viscosity is likely increasing with depth. The net effect may be that the Rayleigh number for the lower mantle is less than $\sim 10^7$.

The linear stability analysis for the onset of convection can also be carried out for a fluid layer heated uniformly from within and cooled from above. The lower boundary is assumed to be insulating, i.e. no heat flows across the boundary. In this case the appropriate Rayleigh number for a fluid layer heated from within is

$$\mathsf{Ra}_H = \frac{\alpha \rho_0^2 g H h^5}{k \eta \kappa}$$

where H is the rate of internal heat generation per unit mass. For no-slip velocity boundary conditions, the minimum critical Rayleigh number is 2772, and the associated value of $2\pi h/\lambda$ is 2.63; for free-slip conditions, the minimum $\mathsf{Ra}_c = 867.8$, and the associated value of $2\pi h/\lambda$ is 1.79.

Additional resources:

- D.L. Turcotte and G. Schubert. *Geodynamics, 3nd edition*. Cambridge University Press, 2014. ISBN: 9780521186230, Section 6.19

- D. Bercovici and G. Schubert. *Treatise on geophysics: Mantle dynamics. Vol. 7*. Elsevier, 2007, Section 2.4.4

- G. Schubert, D.L. Turcotte, and P. Olson. *Mantle Convection in the Earth and Planets*. Cambridge University Press, 2001. ISBN: 0-521-70000-0. DOI: 10.1017/CBO9780511612879, chapter 7

- Pelletier book, chapter 7.2

18.11 Video resources

- *The bizarre patterns that emerge when you heat ANY fluid* by Steve Mould
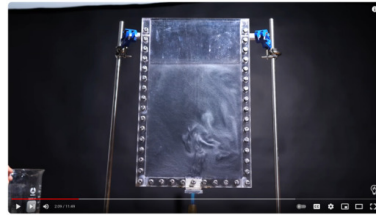
https://youtu.be/kuLX76g7Fec?si=1DSMkXICnr2_yMXv

- *Geodynamics 1: Large-Scale Mantle Convection and Numerical Modeling of it* by Allen McNamara at CIDER 2014

https://www.youtube.com/watch?v=JyHdFKCIYDE
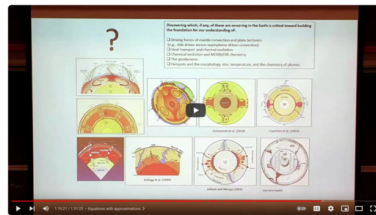
- *Geodynamics/Subduction Lab* by Tobias Hoeink

https://www.youtube.com/watch?v=8cpL5c6lrXA
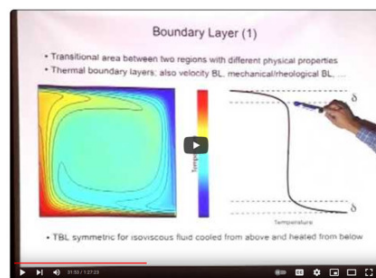
- *Subduction* by Dave stegman
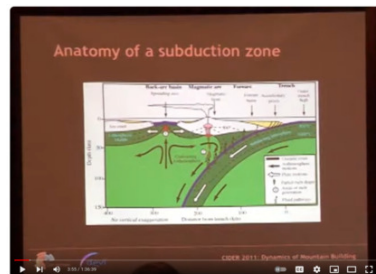
https://youtu.be/f2sJJo8bqqU?si=qBHUOMdVHo8_woNW

18.12 Computer practicals

18.12.1 Introduction

In parameterized convection models for planetary thermal evolution the heat transport characteristics of the convecting mantle are formulated in a pseudo steady state approximation. This is done by parameterization of the surface heatflux as a function of convective vigor, through the non-dimensional Nusselt number Nu. In the convective regime Nu is usually expressed in terms of the Rayleigh number Ra as $\mathsf{Nu} \sim C\,\mathsf{Ra}^\beta$, where C is a constant depending on the domain geometry (please read section 1 of Wolstencroft, Davies, and Davies [1468] (2009) for more information, check also Plumley and Julien [1070] (2019) and references therein, also Korenaga [767] (2003) and Olson [1014] (1987)).

In this lab exercise you will investigate the characteristics of steady-state Rayleigh-Bénard convection and determine the relation between the Nusselt and Rayleigh number experimentally, by means of numerical modelling. In particular you will measure the heatflow through the top surface of a 2D model of a convecting layer, as a function of the Rayleigh number, expressed in the temperature contrast across the convecting layer. This is done by a series of modelling experiments where the coupled equations for thermal convection are solved numerically using finite element methods.

The following sections contain descriptions of the numerical model and the experiments to be done.

18.12.2 Reminder of the governing model equations

In this computerlab you will perform experiments with numerical solutions of the coupled equations describing thermal convection in an incompressible viscous fluid with infinite Prandtl number[27].

In what follows, the assumption is made that geological materials can be treated as fluids (with special properties) within the realm of continuum fluid mechanics and under the Stokes hypothesis. A Boussinesq approximation is applied, neglecting density variations in the equations except in the buoyancy term of the momentum conservation equation. We consider two-dimensional problems.

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \rho(T)\vec{g} = \vec{0} \tag{18.135}$$

$$\vec{\nabla} \cdot \vec{\mathrm{v}} = 0 \tag{18.136}$$

$$\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau} \tag{18.137}$$

$$\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) \tag{18.138}$$

$$\dot{\boldsymbol{\varepsilon}}(\vec{\mathrm{v}}) = \frac{1}{2}\left(\vec{\nabla}\vec{\mathrm{v}} + (\vec{\nabla}\vec{\mathrm{v}})^T\right) \tag{18.139}$$

$$\rho_0 C_p\left(\frac{\partial T}{\partial t} + \vec{\mathrm{v}} \cdot \vec{\nabla}T\right) = \vec{\nabla} \cdot (k\vec{\nabla}T) \tag{18.140}$$

$$\rho(T) = \rho_0(1 - \alpha(T - T_0)) \tag{18.141}$$

Equation (18.135) is the momentum conservation equation and Eq. (18.136) is the mass conservation equation for incompressible fluids. One can resolve the stress tensor $\boldsymbol{\sigma}$ into its spherical part $-p\mathbf{1}$ and its stress deviation $\boldsymbol{\tau}$ (see Eq. (18.137)), where the deviatoric stress tensor is proportional to the strain rate tensor $\dot{\boldsymbol{\varepsilon}}$ (see Eq.(18.138)) through the dynamic viscosity η. Finally Eq. (18.139) relates the strain rate tensor to the velocity field.

---

[27]In heat transfer problems, the Prandtl number controls the relative thickness of the momentum and thermal boundary layers. When Pr is small, it means that the heat diffuses quickly compared to the velocity (momentum).

Equations (18.135), (18.136), (18.137), (18.138) and (18.139) all together lead to the following form of the Stokes equations:

$$\vec{\nabla} \cdot [\eta(\vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\vec{\mathbf{v}}^T)] - \vec{\nabla}p + \rho\vec{g} = \vec{0} \tag{18.142}$$

$$\vec{\nabla} \cdot \vec{\mathbf{v}} = 0 \tag{18.143}$$

Equation (18.142) is an elliptic equation characterized by the fact that changes in buoyancy and constitutive relationships *anywhere* in the domain have an immediate influence on the entire domain.

symbol	meaning and dimension
$\vec{g}$	gravity acceleration vector ($\mathrm{m\,s^{-2}}$)
L_x, L_y	domain size (m)
p	pressure (Pa)
$\boldsymbol{\tau}$	deviatoric stress vector (Pa)
$\vec{\mathbf{v}} = (u, v, w)$	velocity ($\mathrm{m\,s^{-1}}$)
$\dot{\boldsymbol{\varepsilon}}(\vec{\mathbf{v}})$	strain-rate tensor ($\mathrm{s^{-1}}$)
η	viscosity ($\mathrm{Pa\,s}$)
ρ, ρ_0	mass density ($\mathrm{kg\,m^{-3}}$)
$\boldsymbol{\sigma}$	stress tensor (Pa)
k	heat conductivity ($\mathrm{W\,m^{-1}\,K^{-1}}$)
C_p	heat capacity ($\mathrm{J\,K^{-1}}$)
α	thermal expansion ($\mathrm{K^{-1}}$)

18.12.3 Numerical solution of the equations

Introduced in the late 1950s, the finite element method (FEM) [637, 1540, 1541, 1542] has emerged as one of the most powerful numerical methods so far devised.

A thorough mathematical treatment of the finite element formulation of the equations gouverning the physics of the system is beyond the scope of this computer practical, and has been exposed in Section 10 and in various textbooks such as [363] or [537].

We wish to study the system at steady state, which means that it no more changes in time. In practice, only the heat transport equation contains a time (derivative) term so we actually wish to solve these three (coupled) equations:

$$\vec{\nabla} \cdot [\eta(\vec{\nabla}\vec{\mathbf{v}} + \vec{\nabla}\vec{\mathbf{v}}^T)] - \vec{\nabla}p + \rho(T)\vec{g} = \vec{0} \tag{18.144}$$

$$\vec{\nabla} \cdot \vec{\mathbf{v}} = 0 \tag{18.145}$$

$$\rho_0 C_p \vec{\mathbf{v}} \cdot \vec{\nabla}T = k\Delta T \tag{18.146}$$

The main problem is clearly visible in the third one: with respect to the velocity and temperature unknowns this is a nonlinear equation! We therefore have to design a strategy (an algorithm) which will allow us to solve these equations.

One simple approach is as follows: the equations are not solved in a coupled manner. but rather the obtention of a new set of variables $(\boldsymbol{v}, p, T)$ is the product of a two-stage process:

1. assume temperature known, solve for velocity (and pressure) field (the first two equations)

2. assume velocity known, solve for temperature (the last equation).

These two steps need to be repeated as long as the system has not converged to a steady solution. The code therefore implements iterations and these iterations stop when either the maximum number of iterations nstep is reached or convergence has been reached (i.e. temperature and velocity do not change substantially between two consecutive iterations). The structure of the code is then as follows:
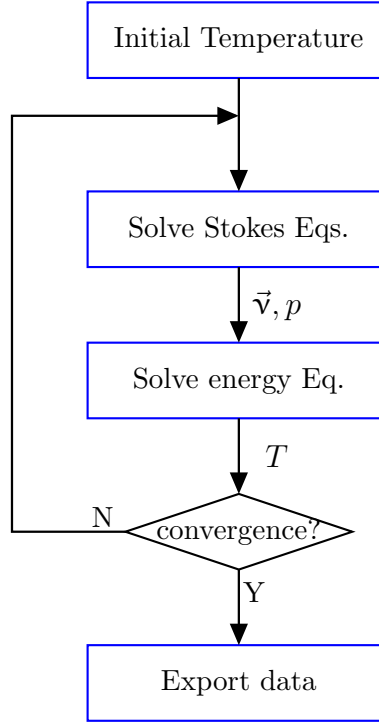
18.12.4 Two-dimensional convection in a unit box

This benchmark deals with the 2-D thermal convection of a fluid of infinite Prandtl number in a rectangular closed cell. In what follows, we will focus on the case 1a, 1b, and 1c experiments as shown in Blankenbach et al. [102] (1989): steady convection with constant viscosity in a square box.

The temperature is fixed to zero on top and to ΔT at the bottom, with reflecting symmetry at the sidewalls (i.e. $\partial_x T = 0$) and there are no internal heat sources. Free-slip conditions are implemented on all boundaries.

The Rayleigh number is given by

$$\mathsf{Ra} = \frac{\alpha g_y \rho_0 \Delta T h^3}{\kappa \eta} = \frac{\alpha g_y \Delta T h^3 \rho^2 C_p}{k \eta} \tag{18.147}$$

In what follows, I use the following parameter values: $L_x = L_y = 1$, $\rho_0 = C_P = k = \eta = 1$, $T_0 = 0$, $\alpha = 10^{-4}$, $g = 10^4 \mathsf{Ra}$.

The initial temperature field is given by

$$T(x, y) = (1 - y) - 0.01 \cos(\pi x / L_x) \sin(\pi y / L_y) \tag{18.148}$$

The perturbation in the initial temperature fields leads to a perturbation of the density field and sets the fluid in motion. Depending on the initial Rayleigh number, the system ultimately reaches a steady state after some time.

The root mean square of the velocity field in the whole domain is defined as follows:

$$\mathbf{v}_{rms} = \left(\frac{1}{V_\Omega} \int_\Omega |\vec{\mathbf{v}}|^2 \, dV \right)^{1/2} = \left(\frac{1}{L_x L_y} \int_\Omega (u^2 + v^2) \, dV \right)^{1/2} \tag{18.149}$$

The Nusselt number (i.e. the mean surface temperature gradient over mean bottom temperature) is computed as follows [102]:

$$\mathsf{Nu} = -L_y \frac{\int_0^{L_x} \frac{\partial T}{\partial y}(y = L_y) dx}{\int_0^{L_x} T(y = 0) dx} \tag{18.150}$$

948

Note that in our case the denominator is equal to L_x since the temperature at the bottom is prescribed to be 1.

Finally, the steady state root mean square velocity v_{rms} and Nusselt number measurements are indicated in the following table alongside those of Blankenbach et al. [102] (1989) and Tackley [1311] (1994). (Note that this benchmark was also carried out and its results published in many other publications [1373, 5, 483, 328, 815] but since they did not provide a complete set of measurement values, they are not included in the table.)

		Blankenbach *et al.* [102]	Tackley [1311]
$Ra = 10^4$	v_{rms}	42.864947 ± 0.000020	42.775
	Nu	4.884409 ± 0.000010	4.878
$Ra = 10^5$	v_{rms}	193.21454 ± 0.00010	193.11
	Nu	10.534095 ± 0.000010	10.531
$Ra = 10^6$	v_{rms}	833.98977 ± 0.00020	833.55
	Nu	21.972465 ± 0.000020	21.998

Steady state Nusselt number Nu and v_{rms} measurements as reported in the literature.

18.12.5 Obtaining the python code

The code is to be downloaded as follows in a terminal:

```
wget https://raw.githubusercontent.com/cedrict/fieldstone/refs/heads/master/python_codes/md/stone
```

If you do not know what a terminal is or if you are using Windows, simply copy

```
https://github.com/cedrict/fieldstone/blob/master/python_codes/md/stone.py
```

in the address bar of your web browser. Then click in the upper right corner on the icon indicating 'Download raw file'. Save the downloaded file in a dedicated folder.

18.12.6 Experiments

To conduct the exercise you can change the following parameters (and run the code until convergence):

- Lx: horizontal extent of the domain (do not change Ly)

- Ra_nb: the Rayleigh number Ra

- tol_ss: the steady state detection tolerance

- nelx,nely: the number of elements in x and y directions

- nstep: the maximum number of iterations to reach steady state

- top_bc_noslip: flag to switch no slip boundary conditions at top boundary (default is free slip)

- bot_bc_noslip: flag to switch no slip boundary conditions at bottom boundary (default is free slip)

Results are written out to *.ascii* files and to *.vtu* files. You can produce plots with python (matplotlib), gnuplot or even excel, as long as these are not pixelated in your report, that they are labeled, captioned, and their axes too. You will find in Appendix L.

Have a thorough look at the code, read **all** instructions, and carry out the following tasks:

1. Determine analytically the expected value of the Nusselt number when there is no convection.

2. Determine the Nusselt number at steady state for a range of Rayleigh numbers, starting from a subcritical value. Produce a plot of Nu against Ra using double logarithmic axes. Determine the critical Rayleigh number.

3. The code produces data files containing 'snapshots' of the resulting numerical solution of the temperature and velocity fields in a suitable format (`.vtu`) for visualization with graphics program *paraview*. Produce colorplots with paraview of the temperature field, for three contrasting Rayleigh number cases, and discuss them.

4. Determine the logarithmic slope or powerlaw index β defined in the introduction.

5. Produce such a Nu − Ra-plot for various grid resolutions. How can you explain the differences in the results ? Produce a plot of Nu as a function of the grid spacing.

6. Look at how the v_{rms} values at steady state depend on Ra.

7. For three contrasting Ra values, plot the temperature profiles (data to be found in *T_profile.ascii*) on a single plot and discuss the obtained figure.

8. Set the number of elements in the horizontal directions to 16. Choose $Ra = 10^5$ and progressively increase the number of points in the vertical direction. Report on the variation of the Nu number at steady state as a function of the vertical resolution.

9. Estimate the value of the critical Rayleigh number from your Nusselt number plot and investigate the difference with the value found in Rayleigh's linear stability analysis for a layer of depth h and infinite horizontal extent, $Ra_C = (27/4)\pi^4$ (see Section 18.10.2).

10. Change the initial temperature profile to something more random, repeat some of these experiments. What can you conclude ?

11. Change the top and bottom boundary conditions from free-slip to no-slip. How does this influence Ra_c ?

12. When convection occurs and a steady state is reached the depth-averaged temperature curves showcases two boundary layers. Measure their thickness as a function of Ra.

13. How high can you set the Rayleigh number and still obtain meaningful results?

14. Bonus: Explore the effect of the aspect ratio of the domain on Ra_c and the slope β.

15. Bonus: change the viscosity function so that the viscosity is 1 in the lower half of the domain and 10^m in the upper half with $m > 1$. Explore & discuss ...

Relevant sources:

- Chapter 3 and the first part of this chapter present the physical equations in more detail.

- Stone 88 shows examples of mantle-scale convection.

- https://youtu.be/YIN9Dcq31x0

- https://youtu.be/5SPCU1sFGGc

- https://youtu.be/ln7QBN0IRTs

- https://youtu.be/d4AS1FmdarU

In what follows the tolerance is set to 10^{-7} and a relaxation factor of 3/4. Results are obtained with `script_onset`. We find that the critical Rayleigh number is between 775 and 780. Also, results are not very sensitive to resolution. When Ra is close to its critical value the scheme requires a lot of iterations as it very slowly converges.

Onset of convection
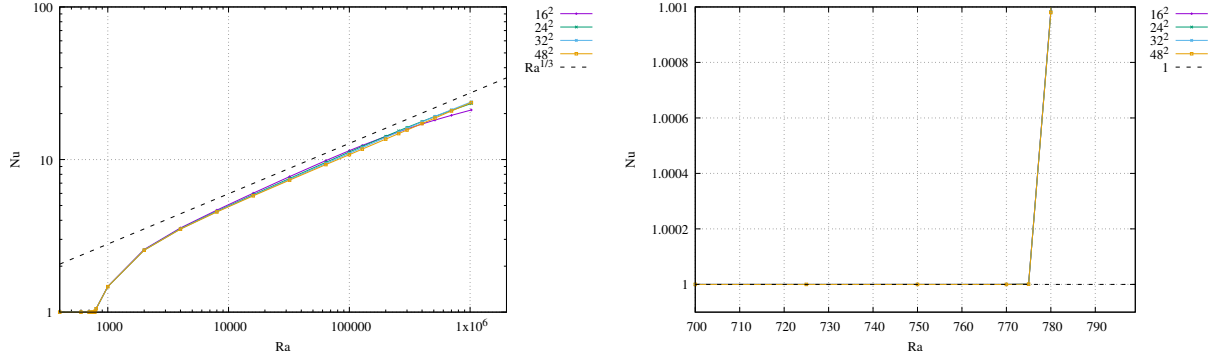
Because the onset of convection (controlled by the critical Rayleigh number seems to be between 770 and 780 we can run a simple test of about 10 iterations only (no steady state is reached) and look at the evolution of the ν_{rms}. If it decreases then convection is dying out and we are in a purely conductive regime. It it increases then convection is going to occur.
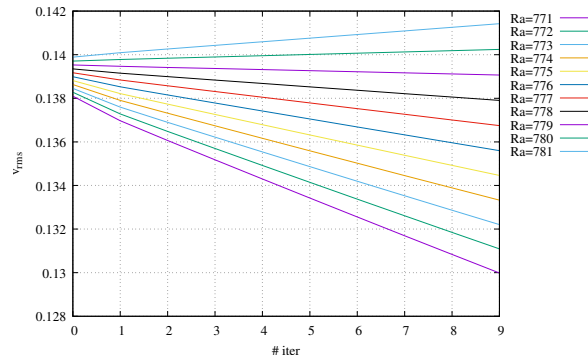
Temperature profiles
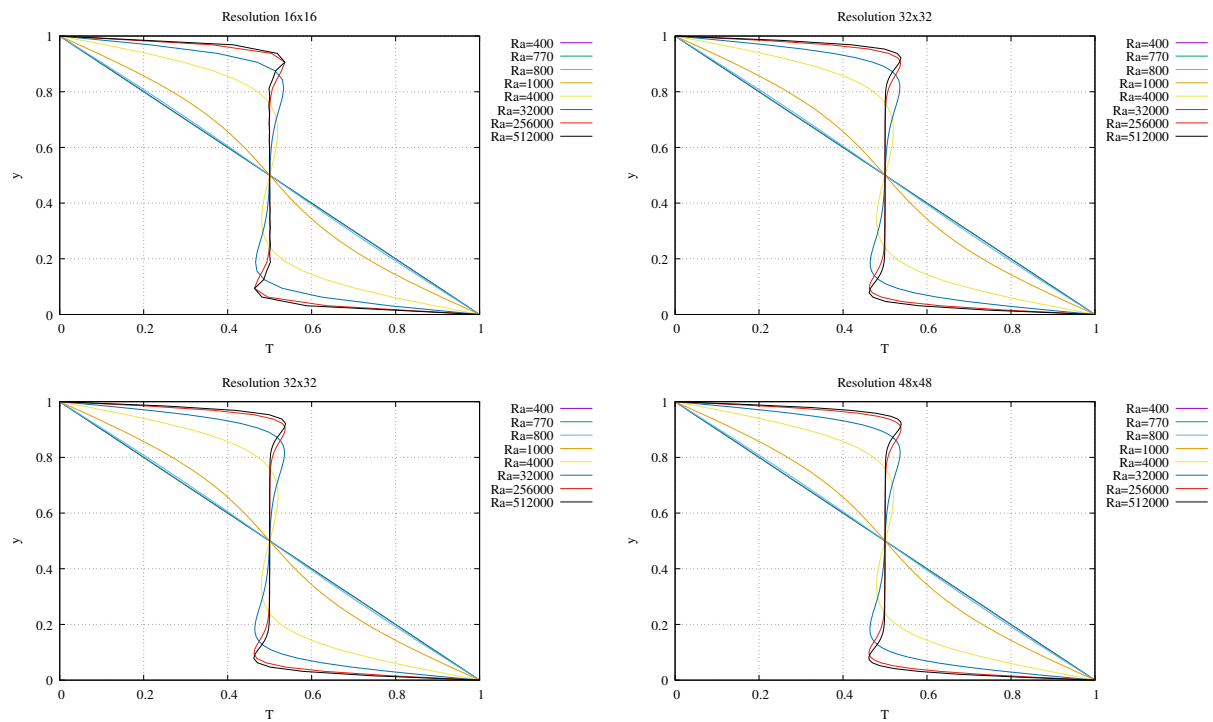
Nusselt number

Results shown for Nusselt number between 700 and 800.
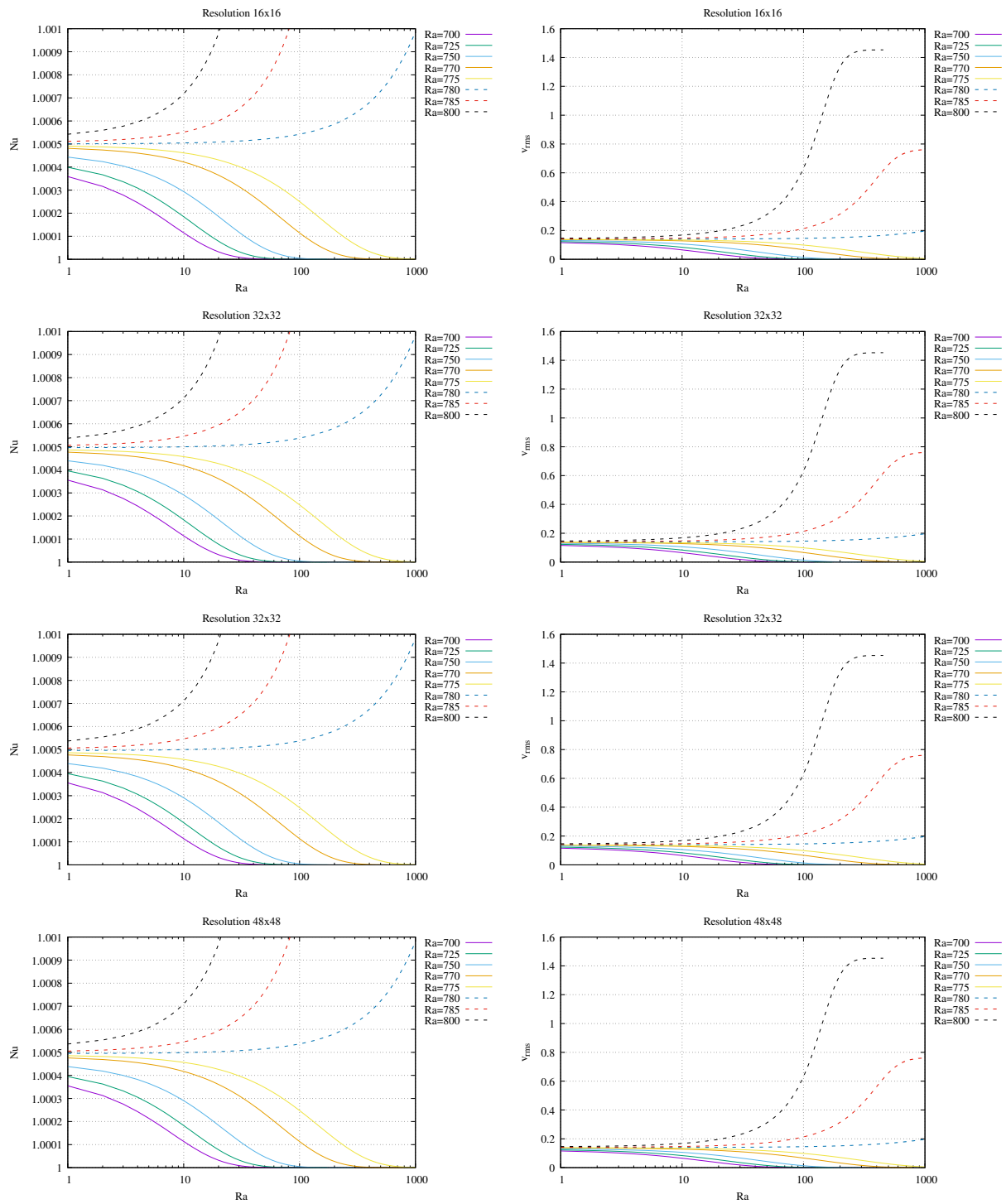
convergence
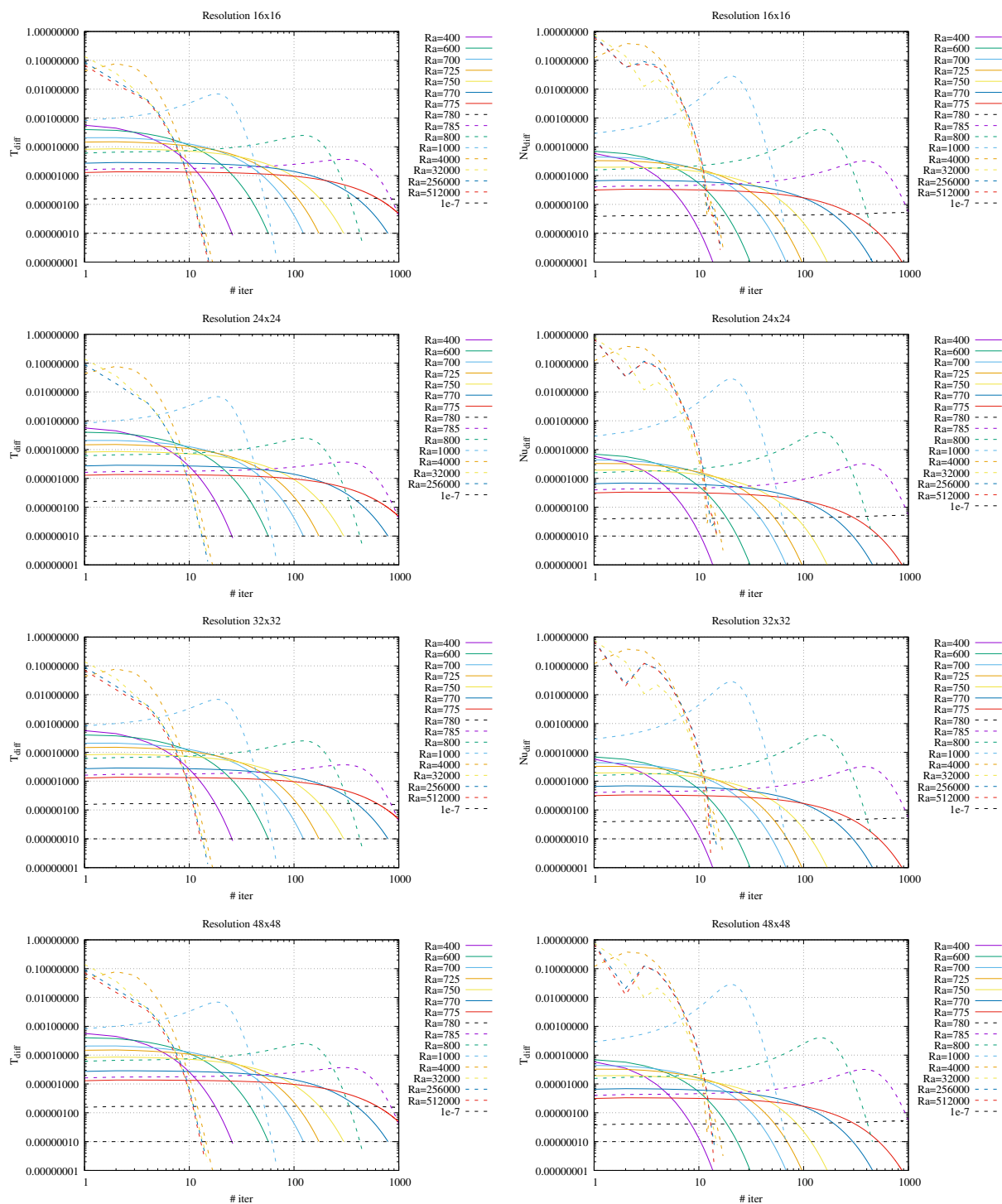
Steady state resolution tests

The following results are obtained with `script_steadystate`. Values are those from Blankenbach et al. [102].
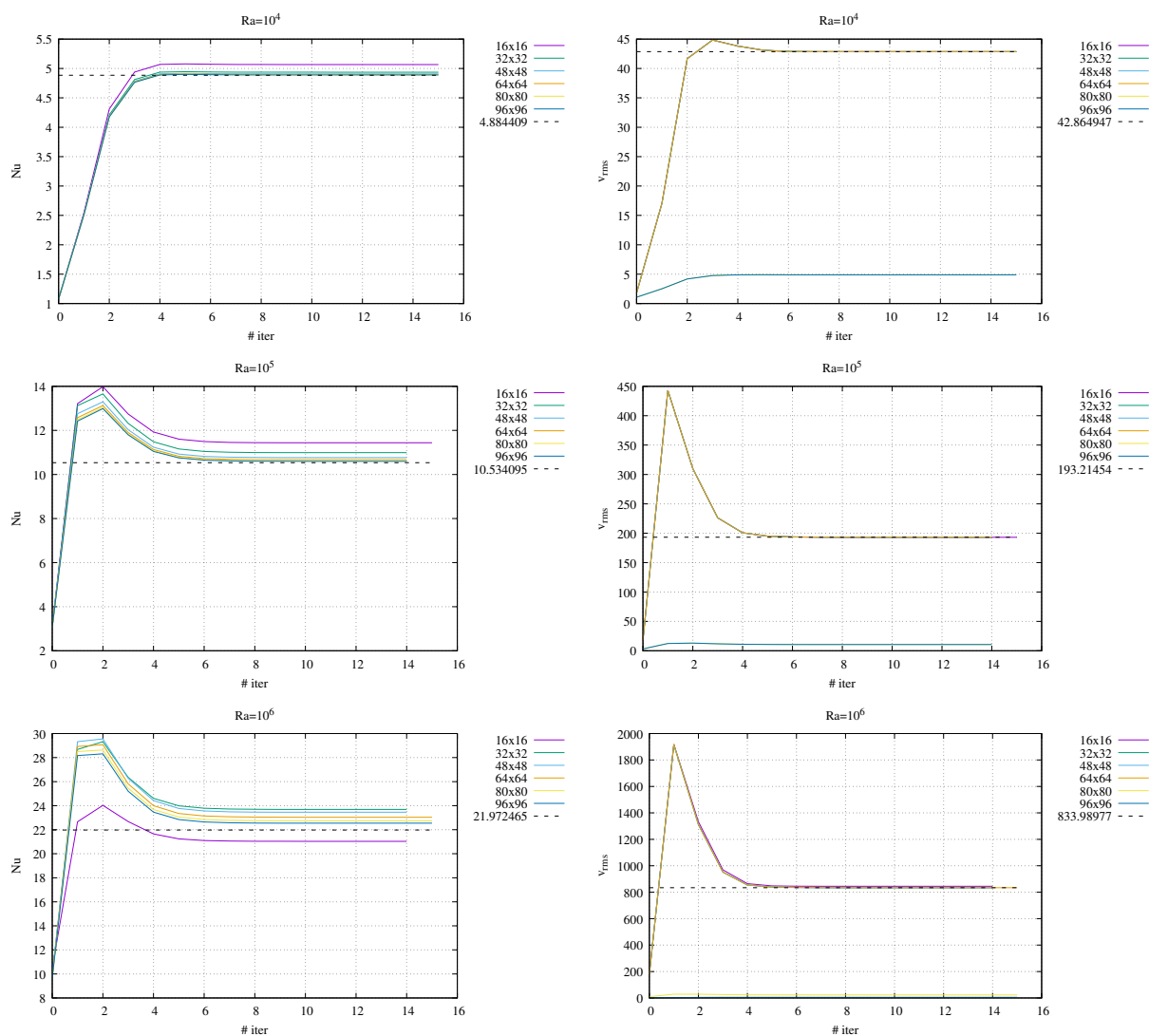
Chapter 19

Manufactured solutions

The method of manufactured solutions is a relatively simple way of carrying out code verification. In essence, one postulates a solution for the PDE at hand (as well as the proper boundary conditions), inserts it in the PDE and computes the corresponding source term. The same source term and boundary conditions will then be used in a numerical simulation so that the computed solution can be compared with the (postulated) true analytical solution.

Examples of this approach are to be found in Donea and Huerta [363], Burstedde et al. [203], Bochev, Dohrmann, and Gunzburger [108], Popov, Lobanov, Popov, Popov, and Gerya [1078], Popov, Lobanov, Popov, and Gerya [1077], Lobanov, Popov, Popov, and Gerya [850], Blinova, Makeev, and Popov [104], Thieulot and Bangerth [1345].

19.1 The repository

I have created in folder `mms` a template python script for incompressible isoviscous isothermal Stokes flow. All one has to do is to provide the velocity and pressure, the strain rate tensor components and its spatial derivatives.

```python
def u_th(x,y):
    return 0

def v_th(x,y):
    return 0

def p_th(x,y):
    return 0

def dpdx_th(x,y):
    return 0

def dpdy_th(x,y):
    return 0

def exx_th(x,y):
    return 0

def exy_th(x,y):
    return 0

def eyy_th(x,y):
    return 0
```

```
def dexxdx(x,y):
    return 0

def dexydx(x,y):
    return 0

def dexydy(x,y):
    return 0

def deyydy(x,y):
    return 0

def bx(x,y):
    return dpdx_th(x,y)-2*dexxdx(x,y)-2*dexydy(x,y)

def by(x,y):
    return dpdy_th(x,y)-2*dexydx(x,y)-2*deyydy(x,y)

def vrms_th():
    return 0

def eta(x,y):
    return 1
```

$$u(x,y) \ =$$
$$v(x,y) \ =$$
$$p(x,y) \ =$$

$$\partial_x u(x,y) \ =$$
$$\partial_y u(x,y) \ =$$
$$\partial_x v(x,y) \ =$$
$$\partial_y v(x,y) \ =$$

$$\dot{\varepsilon}_{xx}(x,y) \ =$$
$$\dot{\varepsilon}_{yy}(x,y) \ =$$
$$\dot{\varepsilon}_{xy}(x,y) \ =$$

$$\frac{\partial p}{\partial x}(x,y) \ =$$
$$\frac{\partial p}{\partial y}(x,y) \ =$$

$$\partial_x \dot{\varepsilon}_{xx}(x,y) \ =$$
$$\partial_x \dot{\varepsilon}_{xy}(x,y) \ =$$
$$\partial_y \dot{\varepsilon}_{xy}(x,y) \ =$$
$$\partial_y \dot{\varepsilon}_{yy}(x,y) \ =$$

958

$$v_{rms} = \sqrt{\frac{1}{L_x L_y} \iint (u^2 + v^2) dx dy} =$$

19.2 Manufactured solution in Donea and Huerta [363] (book)

Taken from [363]. We consider a two-dimensional problem in the square domain $\Omega = [0,1] \times [0,1]$, which possesses a closed-form analytical solution. The problem consists of determining the incompressible flow velocity field $\vec{v} = (u, v)$ and the pressure p such that

$$\vec{\nabla} \cdot (2\eta \dot{\varepsilon}(\vec{v})) - \vec{\nabla} p + \vec{b} = \vec{0} \qquad \text{in } \Omega \tag{19.1}$$

$$\vec{\nabla} \cdot \vec{v} = 0 \qquad \text{in } \Omega \tag{19.2}$$

$$\vec{v} = \vec{0} \qquad \text{on } \Gamma_D \tag{19.3}$$

where the fluid viscosity is taken as $\eta = 1$. The components of the body force $\vec{b}$ are prescribed as

$$
\begin{aligned}
b_x &= (12 - 24y)x^4 + (-24 + 48y)x^3 + (-48y + 72y^2 - 48y^3 + 12)x^2 \\
&\quad + (-2 + 24y - 72y^2 + 48y^3)x + 1 - 4y + 12y^2 - 8y^3 \\
b_y &= (8 - 48y + 48y^2)x^3 + (-12 + 72y - 72y^2)x^2 \\
&\quad + (4 - 24y + 48y^2 - 48y^3 + 24y^4)x - 12y^2 + 24y^3 - 12y^4
\end{aligned}
$$

With this prescribed body force, the exact solution is

$$
\begin{aligned}
u(x,y) &= x^2(1-x)^2(2y - 6y^2 + 4y^3) \\
&= x^2(1-x)^2 2y(1 - 3y + 2y^2) \\
&= x^2(1-x)^2 2y(y-1)(2y-1) \\
v(x,y) &= -y^2(1-y)^2(2x - 6x^2 + 4x^3) \\
&= -y^2(1-y)^2 2x(1 - 3x + 2x^2) \\
&= -y^2(1-y)^2 2x(x-1)(2x-1) \\
p(x,y) &= x(1-x) - 1/6
\end{aligned}
$$

Note that the pressure obeys $\int_\Omega p \, dV = 0$. One can turn to the spatial derivatives of the fields:

$$\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} = (2x - 6x^2 + 4x^3)(2y - 6y^2 + 4y^3) \tag{19.4}$$

$$\dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} = -(2x - 6x^2 + 4x^3)(2y - 6y^2 + 4y^3) \tag{19.5}$$

$$\dot{\varepsilon}_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) = = \frac{1}{2}\left(x^2(1-x)^2(2 - 12y + 12y^2) - y^2(1-y)^2(2 - 12x + 12x^2)\right) \tag{19.6}$$

with of course $\vec{\nabla} \cdot \vec{v} = 0$ and

$$\frac{\partial p}{\partial x} = 1 - 2x \tag{19.7}$$

$$\frac{\partial p}{\partial y} = 0 \tag{19.8}$$

The velocity and pressure fields look like:
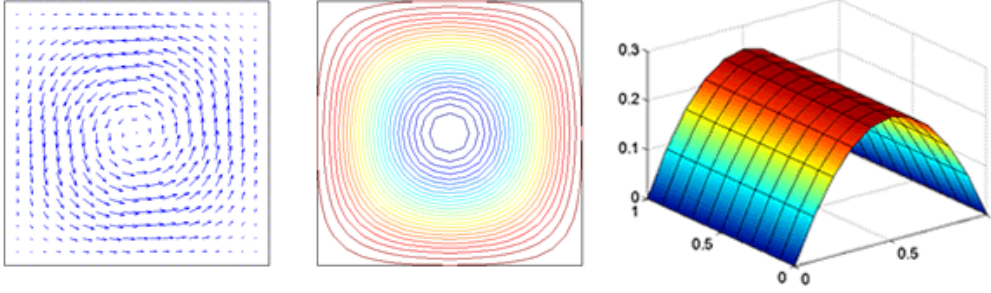
Then the velocity magnitude is given by

$$
\begin{aligned}
|\vec{\mathbf{v}}|(x,y) &= \sqrt{u^2 + v^2} & (19.9)\\
&= \sqrt{[x^2(1-x)^2 2y(1-3y+2y^2)]^2 + [-y^2(1-y)^2 2x(1-3x+2x^2)]^2} & (19.10)\\
&= \sqrt{x^4(1-x)^4 4y^2(1-3y+2y^2)^2 + y^4(1-y)^4 4x^2(1-3x+2x^2)^2} & (19.11)\\
&= \sqrt{4x^2y^2}\sqrt{x^2(1-x)^4(1-3y+2y^2)^2 + y^2(1-y)^4(1-3x+2x^2)^2} & (19.12)\\
&= 2xy\sqrt{x^2(1-x)^4(1-3y+2y^2)^2 + y^2(1-y)^4(1-3x+2x^2)^2} & (19.13)
\end{aligned}
$$

This expression is unfortunately not very useful for later postprocessing...

As shown in [363], If the LBB condition is not satisfied, spurious oscillations spoil the pressure approximation. Figures below show results obtained with a mesh of 20x20 $Q_1 \times P_0$ (left) and $P_1 \times P_1$ (right) elements:
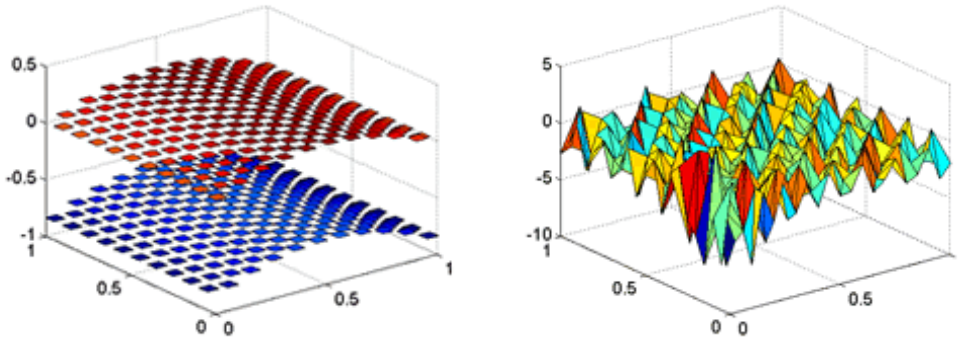
]]

Taking into account that the proposed problem has got analytical solution, it is easy to analyze convergence of the different pairs of elements:
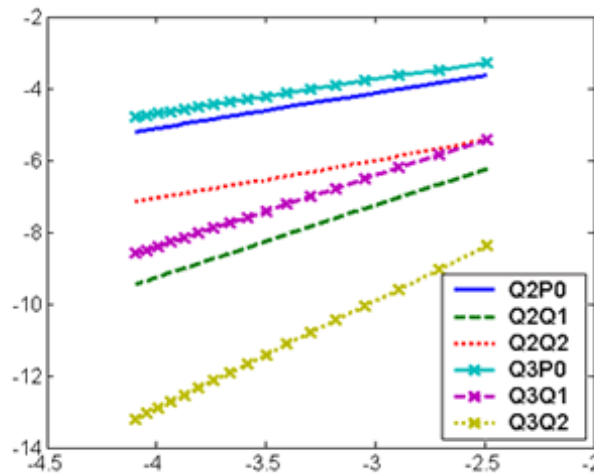
One can also compute the stress components:

$$\sigma_{xx} = 2x^2(2x-2)(4y^3-6y^2+2y) + 4x(-x+1)^2(4y^3-6y^2+2y) - x(-x+1) + 1/6$$
$$\sigma_{xy} = x^2(-x+1)^2(12y^2-12y+2) - y^2(-y+1)^2(12x^2-12x+2)$$
$$\sigma_{yy} = -x(-x+1) - 2y^2(2y-2)(4x^3-6x^2+2x) - 4y(-y+1)^2(4x^3-6x^2+2x) + 1/6$$

All the necessary functions to do this benchmark are in `mms/dh.py`:

```python
# functions for the Donea & Huerta benchmark (dh)

def u_th(x,y):
    return x**2*(1.-x)**2*(2*y-6*y**2+4*y**3)

def v_th(x,y):
    return -y**2*(1.-y)**2*(2*x-6*x**2+4*x**3)

def p_th(x,y):
    return x*(1-x)-1./6.

def dpdx_th(x,y):
    return 1.-2.*x

def dpdy_th(x,y):
    return 0.

def exx_th(x,y):
    return x**2*(2*x-2)*(4*y**3-6*y**2+2*y)+2*x*(-x+1)**2*(4*y**3-6*y**2+2*y)

def exy_th(x,y):
    return (x**2*(-x+1)**2*(12*y**2-12*y+2)-y**2*(-y+1)**2*(12*x**2-12*x+2))/2

def eyy_th(x,y):
    return -exx_th(x,y)

def bx(x,y):
    return ((12.-24.*y)*x**4+(-24.+48.*y)*x*x*x +
            (-48.*y+72.*y*y-48.*y*y*y+12.)*x*x +
            (-2.+24.*y-72.*y*y+48.*y*y*y)*x +
            1.-4.*y+12.*y*y-8.*y*y*y)

def by(x,y):
    return ((8.-48.*y+48.*y*y)*x*x*x*x+
            (-12.+72.*y-72.*y*y)*x*x*x+
            (4.-24.*y+48.*y*y-48.*y*y*y+24.*y**4)*x -
            12.*y*y+24.*y*y*y-12.*y**4)
```

This benchmark is implemented in ASPECT [48] and in STONE 1 and many more.
We have

$$\int_0^1 \int_0^1 u^2 dx dy = \int_0^1 \int_0^1 (x^2(1-x)^2(2y-6y^2+4y^3))^2 dx dy = \frac{1}{33075}$$

$$\int_0^1 \int_0^1 v^2 dx dy = \int_0^1 \int_0^1 (-y^2(1-y)^2(2x-6x^2+4x^3))^2 dx dy = \frac{1}{33075}$$

so the root mean square velocity is

$$v_{rms} = \sqrt{\frac{1}{L_x L_y} \int_0^1 \int_0^1 (u^2+v^2) dx dy} \simeq 0.00777615791$$

We can also look at depth averages. The vertical depth average of the horizontal component of the velocity is given by

$$
\begin{aligned}
\langle u \rangle (y) &= \frac{1}{L_x} \int_0^{L_x} u(x,y) \; dx \\
&= \int_0^{L_x} x^2 (1-x)^2 2y(1-3y+2y^2) \; dx \\
&= \left(\int_0^1 x^2 (1-x)^2 \; dx \right) 2y(1-3y+2y^2) \\
&= \frac{1}{30} 2y(1-3y+2y^2)
\end{aligned}
\tag{19.14}
$$

Likewise, the vertical depth average of the vertical component of the velocity is given by

$$
\begin{aligned}
\langle v \rangle (y) &= \frac{1}{L_x} \int_0^{L_x} v(x,y) \; dx \\
&= -\int_0^1 y^2 (1-y)^2 2x(1-3x+2x^2) \; dx \\
&= -y^2 (1-y)^2 \left(\int_0^1 2x(1-3x+2x^2) \; dx \right) \\
&= 0
\end{aligned}
\tag{19.15}
$$

Unfortunately we have seen in Eq.(19.13) that the velocity magnitude is a rather complex function and we won't be able to compute a depth average analytically.

19.3   SolCx

The SolCx benchmark is intended to test the accuracy of the solution to a problem that has a large jump in the viscosity along a line through the domain. Such situations are common in geophysics: for example, the viscosity in a cold, subducting slab is much larger than in the surrounding, relatively hot mantle material.

The SolCx benchmark computes the Stokes flow field of a fluid driven by spatial density variations, subject to a spatially variable viscosity. Specifically, the domain is $\Omega = [0,1]^2$, gravity is $\vec{g} = (0,-1)^T$ and the density is given by

$$
\rho(x,y) = \sin(\pi y) \cos(\pi x)
\tag{19.16}
$$

Boundary conditions are free slip on all of the sides of the domain and the temperature plays no role in this benchmark. The viscosity is prescribed as follows:

$$
\eta(x,y) = \begin{cases} 1 & \text{for} \quad x < 0.5 \\ 10^6 & \text{for} \quad x > 0.5 \end{cases}
\tag{19.17}
$$

Note the strongly discontinuous viscosity field yields a stagnant flow in the right half of the domain and thereby yields a pressure discontinuity along the interface.

The SolCx benchmark was previously showcased in Duretz *et al.* (2011) [375] and its analytic solution is given in Zhong (1996) [1525]. It has been carried out in Kronbichler *et al.* (2012) [776] and Gerya *et al.* (2013) [480], and is also found in the ASPECT manual [48].

Note that the source code which evaluates the velocity and pressure fields for both SolCx and SolKz is distributed as part of the open source package Underworld (Moresi *et al.* , 2007 [954], `https://www.underworldcode.org/intro-to-underworld/`). I have translated this code to python.
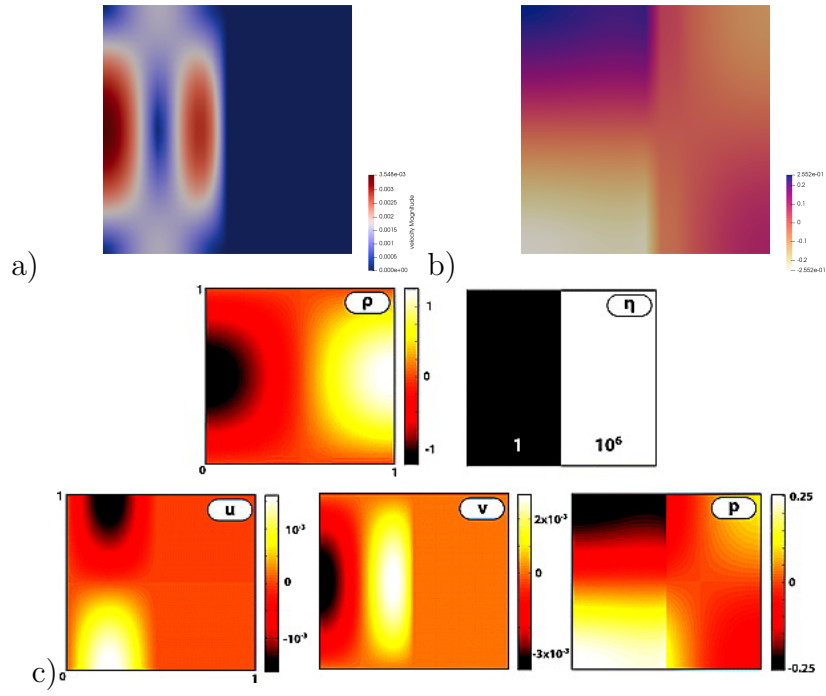
a,b) obtained with ASPECT . c) taken from Duretz *et al.* (2011) [375].
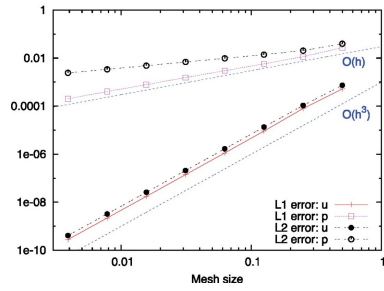
h	$Q_2^d \times Q_1$		$Q_3^d \times Q_2$		$Q_2^d \times P_{-1}$, even mesh		$Q_2^d \times P_{-1}$, odd mesh	
	$\|e_u\|_{L_2}$	$\|e_p\|_{L_2}$	$\|e_u\|_{L_2}$	$\|e_p\|_{L_2}$	$\|e_u\|_{L_2}$	$\|e_p\|_{L_2}$	$\|e_u\|_{L_2}$	$\|e_p\|_{L_2}$
1/8	1.3×10^{-5}	1.4×10^{-2}	6.3×10^{-7}	8.8×10^{-3}	1.3×10^{-5}	1.5×10^{-3}	6.5×10^{-4}	1.1×10^{-2}
1/16	1.7×10^{-6}	9.8×10^{-3}	4.0×10^{-8}	6.2×10^{-3}	1.7×10^{-6}	3.7×10^{-4}	3.6×10^{-4}	6.8×10^{-3}
1/32	2.1×10^{-7}	6.9×10^{-3}	2.6×10^{-9}	4.4×10^{-3}	2.2×10^{-7}	9.2×10^{-5}	1.9×10^{-4}	4.5×10^{-3}
1/64	2.6×10^{-8}	4.9×10^{-3}	1.7×10^{-10}	3.1×10^{-3}	2.6×10^{-8}	2.3×10^{-5}	9.8×10^{-5}	3.3×10^{-3}
1/128	3.3×10^{-9}	3.4×10^{-3}	2.0×10^{-11}	2.2×10^{-3}	3.2×10^{-9}	5.7×10^{-6}	5.0×10^{-5}	2.1×10^{-3}
1/256	4.1×10^{-10}	2.4×10^{-3}	1.7×10^{-11}	1.5×10^{-3}	4.1×10^{-10}	1.4×10^{-6}	2.5×10^{-5}	1.5×10^{-3}
	$\mathcal{O}(h^3)$	$\mathcal{O}(h^{1/2})$	$\mathcal{O}(h^4)$	$\mathcal{O}(h^{1/2})$	$\mathcal{O}(h^3)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h)$	$\mathcal{O}(h^{1/2})$

Taken from Kronbichler *et al.* (2012) [776]. Velocity and pressure errors eu, ep and convergence rates for different choices of the Stokes finite element spaces, using globally refined meshes. For 'odd' meshes, the numbers shown are the average errors from nearby meshes (e.g. for $h = 1/64$, the average of the errors on 63x63 and 65x65 meshes).

📦 Relevant Literature:

- D.A. May and L. Moresi. "Preconditioned iterative methods for Stokes flow problems arising in computational geodynamics". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 33–47. DOI: 10.1016/j.pepi.2008.07.036

- M Velić, L. Moresi, D. May, and M. Knepley. "A Family of Numerically Stable Analytic Solutions for Geodynamic Code Verification". In: ()

- Albert de Montserrat, Jason P Morgan, and Jörg Hasenclever. "LaCoDe: a Lagrangian two-dimensional thermo-mechanical code for large-strain compressible visco-elastic geodynamical modeling". In: *Tectonophysics* 767 (2019), p. 228173. DOI: 10.1016/j.tecto.2019.228173

- Y.A. Mishin, O.V. Vasilyev, and T.V. Gerya. "A Wavelet-Based Adaptive Finite Element Method for the Stokes Problems". In: *Fluids* 7 (2022), p. 221. DOI: 10.3390/fluids7070221

- Ruben Sevilla and Thibault Duretz. "A face-centered finite volume method for high-contrast Stokes interface problems". In: *Int. J. Num. Meth. Eng.* 124 (2023), pp. 3709–3732. DOI: 10.1002/nme.7294

STONE 5, 77

19.4 SolKz

The SolKz benchmark is similar to the SolCx benchmark: the viscosity is a function of the space coordinates too and is given by

$$\eta(y) = \exp(By) \qquad \text{with} \qquad B = 13.8155$$

It is not a discontinuous function but it grows exponentially with the vertical coordinate so that its overall variation is again 10^6. The forcing is chosen by imposing a spatially variable density variation as follows:

$$\rho(x, y) = \sin(2y) \cos(3\pi x)$$

Free slip boundary conditions are imposed on all sides of the domain. This benchmark too is presented in Zhong [1525] (1996) and is studied in Duretz, May, Gerya, and Tackley [375] (2011) and Gerya, May, and Duretz [480] (2013).
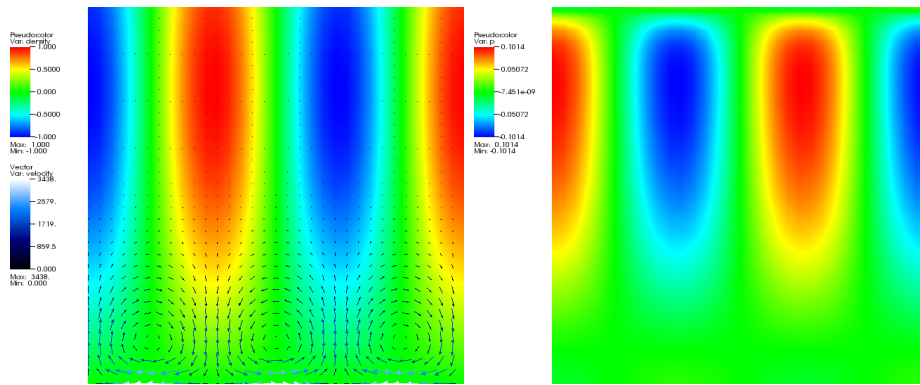
Taken from ASPECT manual [48].
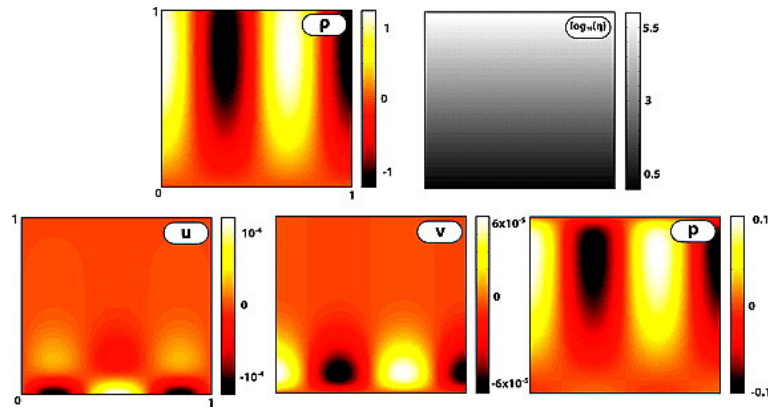
Taken from Duretz *et al.* (2011) [375].

🪨 Relevant Literature: Moresi, Zhong, and Gurnis [955], May and Moresi [895] (2008), [1415], de Montserrat, Morgan, and Hasenclever [341] (2019), Revenaugh and Parsons [1129] (1987), Kronbichler, Heister, and Bangerth [776] (2012). STONE 06, 120.

19.5 SolVi

SolVi is another very common benchmark carried out in the computational geodynamics literature.

This inclusion benchmark solves a problem with a discontinuous viscosity, which is chosen in such a way that the discontinuity is a circle. Given the regular nature of the used by a majority of codes,

this ensures that the discontinuity in the viscosity never aligns to cell boundaries. This in turns leads to almost discontinuous pressures along the interface which are difficult to represent accurately.

Schmid & Podlachikov (2003) [1204]. derived a simple analytic solution for the pressure and velocity fields for such a circular inclusion under simple shear.

A characteristic of the analytical solution is that the pressure is zero inside the inclusion, while outside it follows the relation

$$p_m = 4\dot{\epsilon}\frac{\eta_m(\eta_i - \eta_m)}{\eta_i + \eta_m}\frac{r_i^2}{r^2}\cos(2\theta)$$

where η_i is the viscosity of the inclusion (often taken to be 1000) and $\eta_m 1$ is the viscosity of the background media (often taken to be 1).

One important observation with this benchmark is the fact that the velocity is not zero even far away from the inclusion, so that the analytical solution must be imposed on the sides. Also, because of symmetry, it is often run on the top quadrant $x > 0$, $y > 0$ with free slip imposed on the left and bottom boundaries.
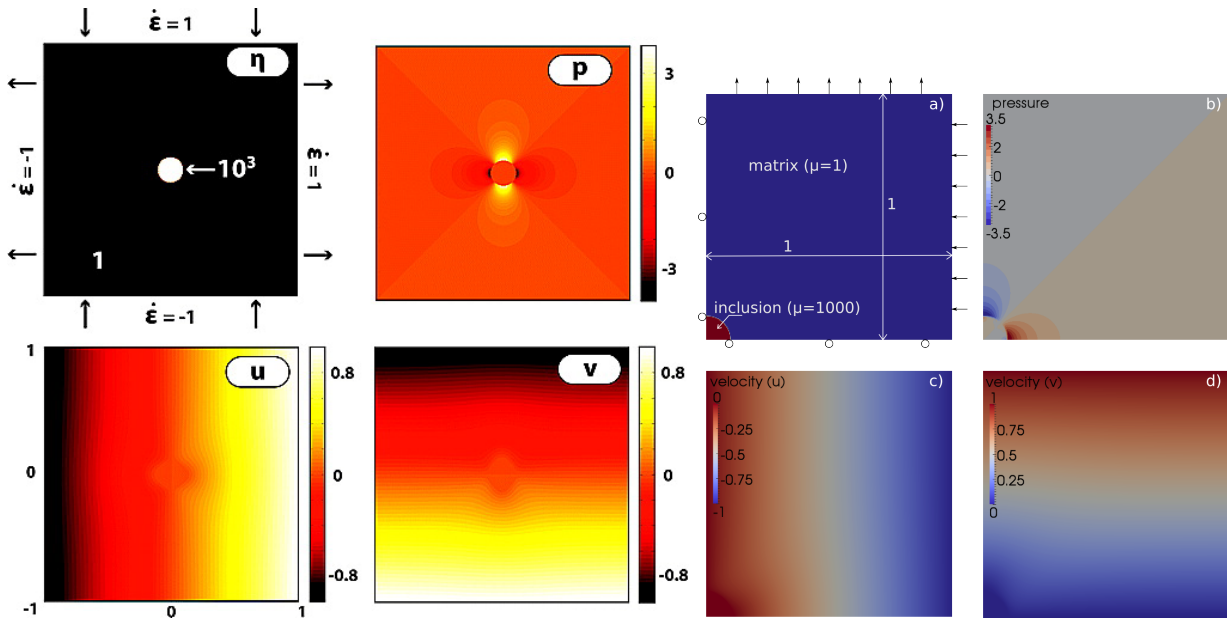
Left: taken from Duretz *et al.* (2011) [375].

📘 Relevant Literature: Kaus and Podlachikov [723], Maierová [871], Deubelbeiss and Kaus [350], Beuchert and Podladchikov [92], Suckale, Nave, and Hager [1302], von Tscharner and Schmalholz [1428], de Montserrat, Morgan, and Hasenclever [341], Bangerth et al. [48], Liu and Tu [845] (2002); Kronbichler, Heister, and Bangerth [776] (2012); Gerya, May, and Duretz [480] (2013); Sevilla and Duretz [1228] (2023). Böhm [117] (2023). STONE 07,

19.6 Manufactured solution with linear pressure

We consider a two-dimensional problem in the square domain $\Omega = [0,1] \times [0,1]$.

$$
\begin{aligned}
u(x,y) &= f(x)g'(y) \\
v(x,y) &= -f'(x)g(y) \\
p(x,y) &= ax + by + c
\end{aligned}
$$

where a, b, c are three real constants and with

$$
\begin{aligned}
f(x) &= x^2(1-x)^2 \\
f'(x) &= 2x - 6x^2 + 4x^3 \\
f''(x) &= 2 - 12x + 12x^2 \\
f'''(x) &= -12 + 24x \\
g(y) &= y^2(1-y)^2 \\
g'(x) &= 2y - 6y^2 + 4y^3 \\
g''(x) &= 2 - 12y + 12y^2 \\
g'''(x) &= -12 + 24y
\end{aligned}
$$

The velocity is the same as in the previous manufactured solution, but the pressure is now linear.
Note that we wish the pressure to be such that $\int_\Omega p \, dV = 0$.

$$
\begin{aligned}
\iint_\Omega p(x,y) dx dy &= \iint_\Omega (ax + by + c) dx dy \\
&= \int_0^1 \int_0^1 ax \, dx dy + \int_0^1 \int_0^1 by \, dx dy + \int_0^1 \int_0^1 c \, dx dy \\
&= a \int_0^1 x \, dx + b \int_0^1 y \, dy + c \\
&= a\frac{1}{2} + b\frac{1}{2} + c
\end{aligned}
$$

which means that we must have $c = -\frac{1}{2}(a+b)$ so that the pressure is given by

$$
p(x,y) = ax + by - \frac{1}{2}(a+b) = a\left(x - \frac{1}{2}\right) + b\left(y - \frac{1}{2}\right)
$$

One can turn to the spatial derivatives of the fields:

$$
\begin{aligned}
\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} &= f'(x)g'(y) \\
\dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} &= -f'(x)g'(y) \\
\dot{\varepsilon}_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) &= \frac{1}{2}(fg'' - f''g)
\end{aligned}
$$

with of course $\vec{\nabla} \cdot \vec{v} = \dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} = 0$. We have

$$
\begin{aligned}
\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} &= f''g' \\
\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} &= \frac{1}{2}(fg''' - f''g') \\
\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} &= \frac{1}{2}(f'g'' - f'''g) \\
\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} &= -f'g''
\end{aligned}
$$

so that the corresponding body force is given by:

$$
\begin{aligned}
b_x &= \frac{\partial p}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} = a - 2f''g' - fg''' + f''g' = a - f''g' - fg''' \\
b_y &= \frac{\partial p}{\partial y} - 2\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} = b + 2f'g'' - f'g'' + f'''g = b + f'g'' + f'''g
\end{aligned}
$$

19.7 Manufactured solution in Dohrmann and Bochev [358] (2004)

Taken from Dohrmann & Bochev (2004,2006) [358, 108]. This benchmark is also used in Worthen *et al.* [1469] and Lamichhane *et al.* [784]. It is for a unit square with $\nu = \eta/\rho = 1$ and the smooth exact solution is

$$
\begin{align}
u(x,y) &= x + x^2 - 2xy + x^3 - 3xy^2 + x^2 y \tag{19.18}\\
v(x,y) &= -y - 2xy + y^2 - 3x^2 y + y^3 - xy^2 \tag{19.19}\\
p(x,y) &= xy + x + y + x^3 y^2 - 4/3 \tag{19.20}
\end{align}
$$

Note that the pressure field is such that $\int_\Omega p \, dV = 0$. The gradient components of the velocity and pressure fields are given by:

$$
\begin{align}
\frac{\partial u}{\partial x} &= 3x^2 + 2x(y+1) - 3y^2 - 2y + 1 \\
\frac{\partial u}{\partial y} &= x(x - 6y - 2) \\
\frac{\partial v}{\partial x} &= -y(6x + y + 2) \\
\frac{\partial v}{\partial y} &= -3x^2 - 2x(y+1) + 3y^2 + 2y - 1 \\
\frac{\partial p}{\partial x} &= 3x^2 y^2 + y + 1 \\
\frac{\partial p}{\partial y} &= 2x^3 y + x + 1
\end{align}
$$

so that the strain rate tensor components are

$$
\begin{align}
\dot{\varepsilon}_{xx} &= 3x^2 + 2x(y+1) - 3y^2 - 2y + 1 \tag{19.21}\\
\dot{\varepsilon}_{yy} &= -3x^2 - 2x(y+1) + 3y^2 + 2y - 1 \tag{19.22}\\
\dot{\varepsilon}_{xy} &= \frac{1}{2}[x(x - 6y - 2) - y(6x + y + 2)] \tag{19.23}\\
&= \frac{1}{2}(x^2 - y^2 - 12xy - 2x - 2y) \tag{19.24}
\end{align}
$$

We have

$$
\begin{align}
\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} &= 2(3x + y + 1) \tag{19.25}\\
\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} &= -6x - y - 1 \tag{19.26}\\
\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} &= -6y + x - 1 \tag{19.27}\\
\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} &= 2(3y - x + 1) \tag{19.28}
\end{align}
$$

so that the corresponding body force is given by:

$$
\begin{aligned}
b_x &= \frac{\partial p}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} \\
&= 3x^2y^2 + y + 1 - 4(3x + y + 1) - 2(-6x - y - 1) \\
&= 3x^2y^2 - y - 1 \\
b_y &= \frac{\partial p}{\partial y} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} \\
&= 2x^3y + x + 1 - 2(-6y + x - 1) - 4(3y - x + 1) \\
&= 2x^3y + 3x - 1
\end{aligned}
$$

(19.29)

(19.30)

Finally we have

$$
\int_0^1 \int_0^1 u^2 dx dy = \int_0^1 \int_0^1 (x + x^2 - 2xy + x^3 - 3xy^2 + x^2y)^2 dx dy = \frac{401}{504}
$$

$$
\int_0^1 \int_0^1 v^2 dx dy = \int_0^1 \int_0^1 (-y - 2xy + y^2 - 3x^2y + y^3 - xy^2)^2 dx dy = \frac{5911}{2520}
$$

so the root mean square velocity is

$$
v_{rms} = \sqrt{\frac{1}{L_x L_y} \int_0^1 \int_0^1 (u^2 + v^2) dx dy} = \sqrt{\frac{401 \cdot 5 + 5911}{2520}} \simeq 1.77236278...
$$
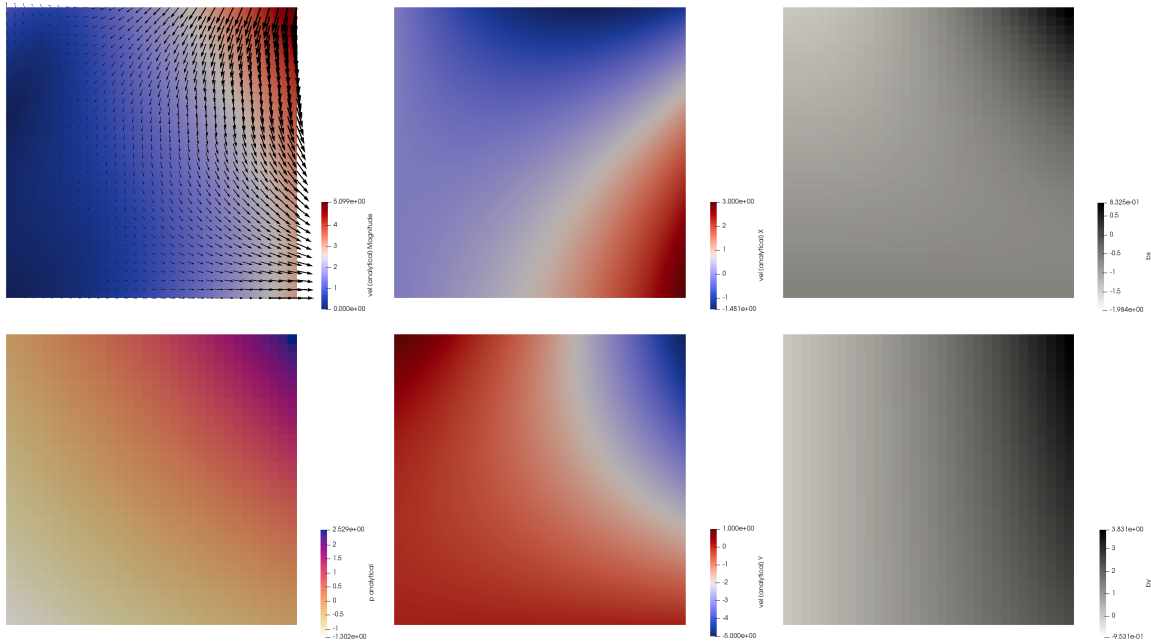
19.8 Manufactured solution in Terrel, Scott, Knepley, Kirby, and Wells [1330] (2012)

This originates in Terrel, Scott, Knepley, Kirby, and Wells [1330] (2012).

$$
\begin{aligned}
u(x,y) &= \sin(4\pi x)\cos(4\pi y) \\
v(x,y) &= -\cos(4\pi x)\sin(4\pi y) \\
p(x,y) &= \pi\cos(4\pi x)\cos(4\pi y) \\
b_x(x,y) &= 28\pi^2\sin(4\pi x)\cos(4\pi y) \\
b_y(x,y) &= -36\pi^2\cos(4\pi x)\sin(4\pi y)
\end{aligned}
\tag{19.31}
$$

$$
\upsilon_{rms} =
$$

19.9 Analytical benchmark - "DB3D"

This benchmark begins by postulating a polynomial solution to the 3D Stokes equation as presented in Dohrmann and Bochev [358] (2004):

$$
\vec{v} = \begin{pmatrix} x + x^2 + xy + x^3 y \\ y + xy + y^2 + x^2 y^2 \\ -2z - 3xz - 3yz - 5x^2 yz \end{pmatrix}
\tag{19.32}
$$

and

$$
p = xyz + x^3 y^3 z - 5/32
\tag{19.33}
$$

While it is then trivial to verify that this velocity field is divergence-free (see here under), the corresponding body force of the Stokes equation can be computed by inserting this solution into the momentum equation with a given viscosity $\eta(x,y,z)$ (constant or position/velocity/strain rate dependent). The domain is a unit cube and velocity boundary conditions simply use Eq. (19.32). Note that the pressure fulfils

$$
\int_\Omega p(x,y,z)dV = 0.
$$

Following Burstedde et al. [203] (2013), the viscosity is given by the smoothly varying function

$$
\eta(x,y,z) = \exp(1 - \beta(x(1-x) + y(1-y) + z(1-z)))
\tag{19.34}
$$

Choosing $\beta = 0$ yields a constant velocity. One can easily show that the ratio of viscosities $\eta^\star$ in the system (the largest viscosity divided by the smallest viscosity) is given by $\eta^\star = \exp(-3\beta/4)$ so that choosing $\beta = 10$ yields $\eta^\star \simeq 1808$ and $\beta = 20$ yields $\eta^\star \simeq 3.269 \times 10^6$. Note that STONE 17 is based on this benchmark.

In order to derive the rhs (i.e. the body force corresponding to the above viscosity, velocity and pressure fields), we start from the momentum conservation equation:

$$
-\vec{\nabla}p + \vec{\nabla} \cdot (2\eta\dot{\boldsymbol{\varepsilon}}(\vec{v})) = \vec{f}
$$

or,

$$
\begin{aligned}
f_x &= -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x}(2\eta\dot{\varepsilon}_{xx}) + \frac{\partial}{\partial y}(2\eta\dot{\varepsilon}_{xy}) + \frac{\partial}{\partial z}(2\eta\dot{\varepsilon}_{xz}) \\
f_x &= -\frac{\partial p}{\partial y} + \frac{\partial}{\partial x}(2\eta\dot{\varepsilon}_{yx}) + \frac{\partial}{\partial y}(2\eta\dot{\varepsilon}_{yy}) + \frac{\partial}{\partial z}(2\eta\dot{\varepsilon}_{yz}) \\
f_x &= -\frac{\partial p}{\partial z} + \frac{\partial}{\partial x}(2\eta\dot{\varepsilon}_{zx}) + \frac{\partial}{\partial y}(2\eta\dot{\varepsilon}_{zy}) + \frac{\partial}{\partial z}(2\eta\dot{\varepsilon}_{zz})
\end{aligned}
$$

Let us compute all the strain rate components:

$$
\begin{aligned}
\dot{\varepsilon}_{xx} &= 1 + 2x + y + 3x^2 y \\
\dot{\varepsilon}_{yy} &= 1 + x + 2y + 2x^2 y \\
\dot{\varepsilon}_{zz} &= -2 - 3x - 3y - 5x^2 y \\
2\dot{\varepsilon}_{xy} &= (x + x^3) + (y + 2xy^2) = x + y + 2xy^2 + x^3 \\
2\dot{\varepsilon}_{xz} &= (0) + (-3z - 10xyz) = -3z - 10xyz \\
2\dot{\varepsilon}_{yz} &= (0) + (-3z - 5x^2 z) = -3z - 5x^2 z
\end{aligned}
$$

In passing, one can easily verify that $\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} = 0$. We further have

$$
\begin{aligned}
\frac{\partial}{\partial x} 2\dot{\varepsilon}_{xx} &= 2(2 + 6xy) \\
\frac{\partial}{\partial y} 2\dot{\varepsilon}_{xy} &= 1 + 4xy \\
\frac{\partial}{\partial z} 2\dot{\varepsilon}_{xz} &= -3 - 10xy \\
\frac{\partial}{\partial x} 2\dot{\varepsilon}_{xy} &= 1 + 2y^2 + 3x^2 \\
\frac{\partial}{\partial y} 2\dot{\varepsilon}_{yy} &= 2(2 + 2x^2) \\
\frac{\partial}{\partial z} 2\dot{\varepsilon}_{yz} &= -3 - 5x^2 \\
\frac{\partial}{\partial x} 2\dot{\varepsilon}_{xz} &= -10yz \\
\frac{\partial}{\partial y} 2\dot{\varepsilon}_{yz} &= 0 \\
\frac{\partial}{\partial z} 2\dot{\varepsilon}_{zz} &= 2(0)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial p}{\partial x} &= yz + 3x^2 y^3 z \\
\frac{\partial p}{\partial y} &= xz + 3x^3 y^2 z \\
\frac{\partial p}{\partial z} &= xy + x^3 y^3
\end{aligned}
$$

Constant viscosity case

Choosing $\beta = 0$ yields a constant velocity $\eta(x, y, z) = \exp(1) \simeq 2.718$ (and greatly simplifies the right-hand side) so that

$$
\frac{\partial \eta}{\partial x} = \frac{\partial \eta}{\partial y} = \frac{\partial \eta}{\partial z} = 0
$$

and

$$
\begin{aligned}
f_x &= -\frac{\partial p}{\partial x} + 2\eta\frac{\partial}{\partial x}\dot{\varepsilon}_{xx} + 2\eta\frac{\partial}{\partial y}\dot{\varepsilon}_{xy} + 2\eta\frac{\partial}{\partial z}\dot{\varepsilon}_{xz} \\
&= -(yz + 3x^2y^3z) + 2(2 + 6xy) + (1 + 4xy) + (-3 - 10xy) \\
&= -(yz + 3x^2y^3z) + \eta(2 + 6xy) \\
f_y &= -\frac{\partial p}{\partial y} + 2\eta\frac{\partial}{\partial x}\dot{\epsilon}_{xy} + 2\eta\frac{\partial}{\partial y}\dot{\epsilon}_{yy} + 2\eta\frac{\partial}{\partial z}\dot{\epsilon}_{yz} \\
&= -(xz + 3x^3y^2z) + \eta(1 + 2y^2 + 3x^2) + \eta 2(2 + 2x^2) + \eta(-3 - 5x^2) \\
&= -(xz + 3x^3y^2z) + \eta(2 + 2x^2 + 2y^2) \\
f_z &= -\frac{\partial p}{\partial z} + 2\eta\frac{\partial}{\partial x}\dot{\epsilon}_{xz} + 2\eta\frac{\partial}{\partial y}\dot{\epsilon}_{yz} + 2\eta\frac{\partial}{\partial z}\dot{\epsilon}_{zz} \\
&= -(xy + x^3y^3) + \eta(-10yz) + 0 + 0 \\
&= -(xy + x^3y^3) + \eta(-10yz)
\end{aligned}
$$

Finally

$$
\vec{f} = -\begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \eta\begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix} \tag{19.35}
$$

Remark. *There seems to be a sign problem with Eq.(26) in Burstedde et al. [203]:*

$$
\begin{aligned}
\mathbf{f} = -\nabla p + \mu &\begin{pmatrix} -2 - 6xy \\ -2 - 2x^2 - 2y^2 \\ 10yz \end{pmatrix} \\
-\mu_x &\begin{pmatrix} 2 + 4x + 2y + 6xy \\ x + x^3 + y + 2xy^2 \\ -3z - 10xyz \end{pmatrix} - \mu_y\begin{pmatrix} x + x^3 + y + 2xy^2 \\ 2 + 2x + 4y + 4x^2y \\ -3z - 5x^2z \end{pmatrix} \\
-\mu_z &\begin{pmatrix} -3z - 10xyz \\ -3z - 5x^2z \\ -4 - 6x - 6y - 10x^2y \end{pmatrix}.
\end{aligned} \tag{26}
$$

19.9.1 Variable viscosity case

The spatial derivatives of the viscosity are then given by

$$
\begin{aligned}
\frac{\partial}{\partial x}\eta(x,y,z) &= -(1 - 2x)\beta\,\eta(x,y,z) \\
\frac{\partial}{\partial y}\eta(x,y,z) &= -(1 - 2y)\beta\,\eta(x,y,z) \\
\frac{\partial}{\partial z}\eta(x,y,z) &= -(1 - 2z)\beta\,\eta(x,y,z)
\end{aligned}
$$

and the right-hand side by

$$
\begin{aligned}
\vec{f} &= -\begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \eta \begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix} - (1-2x)\beta\eta(x,y,z) \begin{pmatrix} 2\dot{\epsilon}_{xx} \\ 2\dot{\epsilon}_{xy} \\ 2\dot{\epsilon}_{xz} \end{pmatrix} \\
&\quad -(1-2y)\beta\eta(x,y,z) \begin{pmatrix} 2\dot{\epsilon}_{xy} \\ 2\dot{\epsilon}_{yy} \\ 2\dot{\epsilon}_{yz} \end{pmatrix} - (1-2z)\beta\eta(x,y,z) \begin{pmatrix} 2\dot{\epsilon}_{xz} \\ 2\dot{\epsilon}_{yz} \\ 2\dot{\epsilon}_{zz} \end{pmatrix} \\
&= -\begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \eta \begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix} - (1-2x)\beta\eta \begin{pmatrix} 2 + 4x + 2y + 6x^2y \\ x + y + 2xy^2 + x^3 \\ -3z - 10xyz \end{pmatrix} \\
&\quad -(1-2y)\beta\eta \begin{pmatrix} x + y + 2xy^2 + x^3 \\ 2 + 2x + 4y + 4x^2y \\ -3z - 5x^2z \end{pmatrix} - (1-2z)\beta\eta \begin{pmatrix} -3z - 10xyz \\ -3z - 5x^2z \\ -4 - 6x - 6y - 10x^2y \end{pmatrix}
\end{aligned}
\tag{19.36}
$$

which translates as follows in Python:

```python
def bx(x,y,z,beta):
    mu=np.exp(1-beta*(x*(1-x)+y*(1-y)+z*(1-z)) )
    mux=-beta*(1-2*x)*mu
    muy=-beta*(1-2*y)*mu
    muz=-beta*(1-2*z)*mu
    val=-(y*z+3*x**2*y**3*z) + mu * (2+6*x*y) \
        +(2+4*x+2*y+6*x**2*y) * mux \
        +(x+x**3+y+2*x*y**2 ) * muy \
        +(-3*z-10*x*y*z      ) * muz
    return val

def by(x,y,z,beta):
    mu=np.exp(1-beta*(x*(1-x)+y*(1-y)+z*(1-z)) )
    mux=-beta*(1-2*x)*mu
    muy=-beta*(1-2*y)*mu
    muz=-beta*(1-2*z)*mu
    val=-(x*z+3*x**3*y**2*z) + mu * (2 +2*x**2 + 2*y**2) \
        +(x+x**3+y+2*x*y**2    ) * mux \
        +(2+2*x+4*y+4*x**2*y   ) * muy \
        +(-3*z-5*x**2*z        ) * muz
    return val

def bz(x,y,z,beta):
    mu=np.exp(1-beta*(x*(1-x)+y*(1-y)+z*(1-z)) )
    mux=-beta*(1-2*x)*mu
    muy=-beta*(1-2*y)*mu
    muz=-beta*(1-2*z)*mu
    val=-(x*y+x**3*y**3) + mu * (-10*y*z) \
        +(-3*z-10*x*y*z           ) * mux \
        +(-3*z-5*x**2*z           ) * muy \
        +(-4-6*x-6*y-10*x**2*y ) * muz
    return val
```

19.9.2 root mean square velocity

Let us now compute the root mean square velocity:

$$
\begin{aligned}
\int_\Omega u^2 dxdydz &= \int_0^{+1}\int_0^{+1}\int_0^{+1} (x + x^2 + xy + x^3 y)^2 dxdydz = 2867/1260 \\
\int_\Omega v^2 dxdydz &= \int_0^{+1}\int_0^{+1}\int_0^{+1} (y + xy + y^2 + x^2 y^2)^2 dxdydz = 3947/1800 \\
\int_\Omega w^2 dxdydz &= \int_0^{+1}\int_0^{+1}\int_0^{+1} (-2z - 3xz - 3yz - 5x^2 yz)^2 dxdydz = 463/36
\end{aligned}
$$

then

$$
\boldsymbol{\upnu}_{rms} = \sqrt{2867/1260 + 3947/1800 + 463/36} \simeq 4.1628459
$$

19.10   Analytical benchmark in Bercovier and Engelman [84] (1979)

From [84]. The two-dimensional domain is a unit square. The body forces are:

$$
\begin{aligned}
f_x &= 128[x^2(x-1)^2 12(2y-1) + 2(y-1)(2y-1)y(12x^2 - 12x + 2)] \\
f_y &= 128[y^2(y-1)^2 12(2x-1) + 2(x-1)(2x-1)y(12y^2 - 12y + 2)]
\end{aligned}
\tag{19.37}
$$

The solution is

$$
\begin{aligned}
u &= -256x^2(x-1)^2 y(y-1)(2y-1) \\
v &= 256y^2(y-1)^2 x(x-1)(2x-1) \\
p &= 0
\end{aligned}
\tag{19.38}
$$

$$
\begin{aligned}
du/dx &= 512(1-2x)(-1+x)x(-1+y)y(-1+2y) & (19.39)\\
du/dy &= -256(-1+x)^2 x^2(1-6y+6y^2) & (19.40)\\
dv/dx &= 256y^2(y-1)^2 x(x-1)(2x-1) & (19.41)\\
dv/dy &= -512(-1+x)x(1-2x)(-1+y)y(-1+2y) & (19.42)\\
 & & (19.43)
\end{aligned}
$$

and we can easily verify that $\vec{\nabla}\cdot\vec{v} = du/dx + dv/dy = 0$.
CHECK RHS !
Another choice with a non-zero pressure:

$$
\begin{aligned}
f_x &= 128[x^2(x-1)^2 12(2y-1) + 2(y-1)(2y-1)y(12x^2 - 12x + 2)] + y - 1/2 \\
f_y &= 128[y^2(y-1)^2 12(2x-1) + 2(x-1)(2x-1)y(12y^2 - 12y + 2)] + x - 1/2
\end{aligned}
\tag{19.44}
$$

The solution is

$$
\begin{aligned}
u &= -256x^2(x-1)^2 y(y-1)(2y-1) \\
v &= 256y^2(y-1)^2 x(x-1)(2x-1) \\
p &= (x-1/2)(y-1/2)
\end{aligned}
\tag{19.45}
$$

19.11 Analytical benchmark in Ilinca and Pelletier [653] (2007)

This is taken from [653].

Let us consider the Poiseuille flow of a Newtonian fluid. The channel has isothermal flat walls located at $y = \pm h$. The velocity distribution is parabolic:

$$u = u_0 \left(1 - \frac{y^2}{h^2} \right) \qquad v = 0$$

where u_0 is the maximum velocity. The (steady state) temperature field is the solution of the advection-diffusion equation:

$$\rho C_p \vec{v} \cdot \vec{\nabla} T = k \Delta T + \Phi$$

where Φ is the dissipation function given by

$$\Phi = \eta \left[2 \left(\frac{\partial u}{\partial x} \right)^2 + 2 \left(\frac{\partial v}{\partial y} \right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right)^2 \right] = \eta \left(\frac{\partial u}{\partial y} \right)^2 = 4\eta \frac{u_0^2 y^2}{h^4}$$

We logically assume that $T = T(y)$ so that $\partial T / \partial x = 0$ and $\vec{v} \cdot \vec{\nabla} T = 0$. We then have to solve:

$$k \frac{\partial^2 T}{\partial y^2} + 4\eta \frac{u_0^2 y^2}{h^4} = 0$$

We can integrate twice and use the boundary conditions $T(y = \pm h) = T_0$ to arrive at:

$$T(y) = T_0 + \frac{1}{3} \frac{\eta u_0^2}{k} \left[1 - \left(\frac{y}{h} \right)^4 \right]$$

with a maximum temperature

$$T_M = T(y = 0) = T_0 + \frac{1}{3} \frac{\eta u_0^2}{k}$$

19.12 Analytical benchmark - "grooves"

This benchmark was designed by Dave May. The velocity and pressure fields are given by

$$
\begin{aligned}
u(x,y) &= x^3 y + x^2 + xy + x \\
v(x,y) &= -\frac{3}{2} x^2 y^2 - 2xy - \frac{1}{2} y^2 - y \\
p(x,y) &= x^2 y^2 + xy + 5 + p_0
\end{aligned}
\tag{19.46}
$$

where p_0 is a constant to be determined based on the type of pressure normalisation. The viscosity is chosen to be

$$\eta(x,y) = -\sin(p) + 1 + \epsilon = -\sin(x^2 y^2 + xy + 5) + 1 + \epsilon \tag{19.47}$$

974

where ϵ actually controls the viscosity contrast. Note that inserting the polynomial expression of the pressure inside the viscosity expression makes the problem linear. We have

$$
\begin{aligned}
\dot\varepsilon_{xx} = \frac{\partial u}{\partial x} &= 3x^2y + 2x + y + 1 \\
\dot\varepsilon_{yy} = \frac{\partial v}{\partial y} &= -3x^2y - 2x - y - 1 \\
\dot\varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) &= \frac{1}{2}\left(x^3 + x - 3xy^2 - 2y\right)
\end{aligned} \tag{19.48}
$$

and we can verify that the velocity field is incompressible since $\vec\nabla \cdot \vec{v} = \dot\varepsilon_{xx} + \dot\varepsilon_{yy} = 0$. The pressure gradient is given by

$$
\begin{aligned}
\frac{\partial p}{\partial x} &= 2xy^2 + y \\
\frac{\partial p}{\partial y} &= 2x^2y + x
\end{aligned}
$$

The right hand side term of the Stokes equation is such that

$$
\begin{aligned}
-\frac{\partial p}{\partial x} + \frac{\partial s_{xx}}{\partial x} + \frac{\partial s_{yx}}{\partial y} + f_x &= 0 \\
-\frac{\partial p}{\partial y} + \frac{\partial s_{xy}}{\partial x} + \frac{\partial s_{yy}}{\partial y} + f_y &= 0
\end{aligned} \tag{19.49}
$$

with

$$
\begin{aligned}
\frac{\partial s_{xx}}{\partial x} &= \frac{\partial(2\eta\dot\varepsilon_{xx})}{\partial x} = 2\eta\frac{\partial\dot\varepsilon_{xx}}{\partial x} + 2\frac{\partial\eta}{\partial x}\dot\varepsilon_{xx} \\
\frac{\partial s_{zx}}{\partial z} &= \frac{\partial(2\eta\dot\varepsilon_{zx})}{\partial z} = 2\eta\frac{\partial\dot\varepsilon_{zx}}{\partial z} + 2\frac{\partial\eta}{\partial z}\dot\varepsilon_{zx} \\
\frac{\partial s_{xz}}{\partial x} &= \frac{\partial(2\eta\dot\varepsilon_{xz})}{\partial x} = 2\eta\frac{\partial\dot\varepsilon_{xz}}{\partial x} + 2\frac{\partial\eta}{\partial x}\dot\varepsilon_{xz} \\
\frac{\partial s_{zz}}{\partial z} &= \frac{\partial(2\eta\dot\varepsilon_{zz})}{\partial z} = 2\eta\frac{\partial\dot\varepsilon_{zz}}{\partial z} + 2\frac{\partial\eta}{\partial z}\dot\varepsilon_{zz} \\
\frac{\partial\eta}{\partial x} &= -z(2xz + 1)\cos(x^2z^2 + xz + 5) \\
\frac{\partial\eta}{\partial z} &= -x(2xz + 1)\cos(x^2z^2 + xz + 5) \\
\frac{\partial\dot\varepsilon_{xx}}{\partial x} &= 6xz + 2 \\
\frac{\partial\dot\varepsilon_{zx}}{\partial z} &= -3xz - 1 \\
\frac{\partial\dot\varepsilon_{xz}}{\partial x} &= \frac{1}{2}(3x^2 + 1 - 3z^2) \\
\frac{\partial\dot\varepsilon_{zz}}{\partial z} &= -3x^2 - 1
\end{aligned}
$$

Velocity boundary conditions are prescribed on all four boundaries so that the pressure is known up to a constant (the pressure solution has a nullspace), and the p_0 constant can be determined by requiring that

$$
\int_0^L \int_0^L p(x,y)\, dxdy = \int_0^L \int_0^L (x^2y^2 + xy + 5)dxdy + \int_0^L \int_0^L p_0\, dxdy = \int_0^L \int_0^L (x^2y^2 + xy + 5)dxdy + p_0 L^2 = 0
$$

where L is the size of the square domain. Then

$$p_0 = -\frac{1}{L^2} \int_0^L \int_0^L (x^2 y^2 + xy + 5) dx dy = -\frac{L^4}{9} - \frac{L^2}{4} - 5$$

As seen in the following figure, the value of ϵ controls the viscosity field amplitude. This is simply explained by the fact that when the sin term of the viscosity takes value 1, the viscosity is then equal to ϵ.
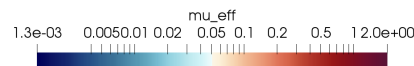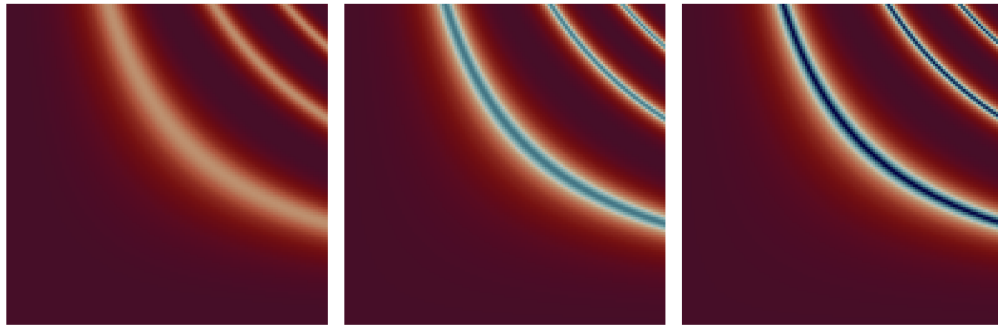
Domain size 2x2 with $\epsilon = 0.1, 0.01, 0.001$

Another interesting aspect of this benchmark is the fact that increasing the domain size adds complexity to it as it increases the number of low viscosity zones and the spacing between them also decreases:
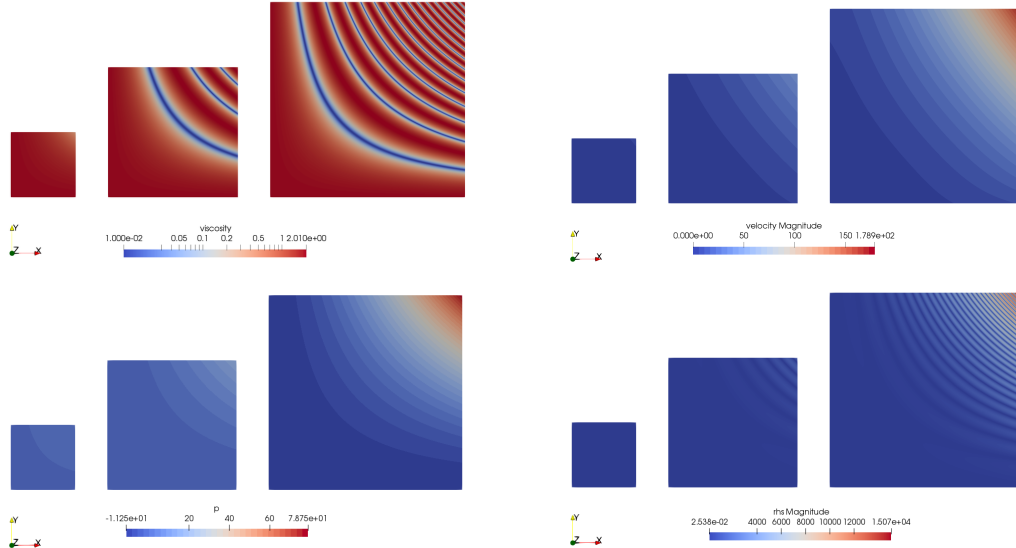
Three different domain sizes (1x1, 2x2, 3x3) with $\epsilon = 0.001$.

Finally, because the analytical expression for both components of the velocity is a polynomial, we can also compute the root mean square velocity exactly. For instance, for a 2x2 domain:
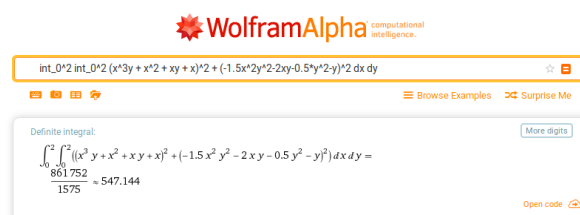
and we end up with (for $L = 2$)

$$v_{rms} = \sqrt{\frac{1}{L^2} \frac{861752}{1575}} = \sqrt{\frac{215438}{1575}} \simeq 11.6955560683$$

I have added this benchmark to ASPECT . The velocity and pressure errors (in the L_2 norm) are measured for $L = 1, 2, 3$, levels 3 to 9 (resolutions 8×8 to 512×512) and $\epsilon = 10^{-1}, 10^{-2}, 10^{-3}$. The figure below shows the velocity and pressure error convergence as a function of the mesh size for $\epsilon = 0.1$ (results are identical for the other two ϵ values). The expected convergence rates (cubic convergence for velocity and quadratic for pressure) are recovered for the 1×1 domain at all resolutions. These rates are recovered for the 2×2 domain for resolutions above level 6. We see that the multitude of low viscosity bands in the upper right corner of the 3×3 domain will require a refinement level superior to 9 to recover the optimal convergence rates.
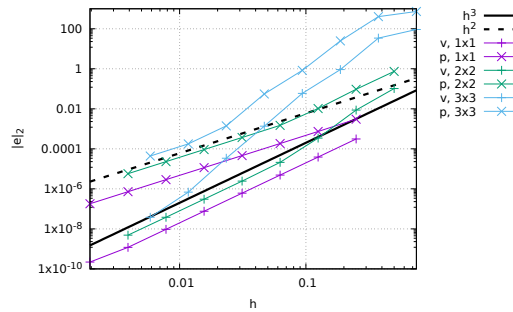
Velocity and pressure error convergence as a function of the mesh size h for 3 domain sizes with $\epsilon = 0.1$.

This benchmark is implemented and used in STONE 112.

19.13 Analytical benchmark - "Kovasznay"

This flow was published by L.I.G. Kovasznay in 1948 [770]. This paper presents an exact two-dimensional solution of the Navier-Stokes equations with a periodicity in the vertical direction, gives an analytical solution to the steady-state Navier-Stokes equations that is similar which is a flow-field behind a periodic array of cylinders.

$$u(x, y) = 1 - \exp(\lambda x) \cos(2\pi y) \tag{19.50}$$

$$v(x, y) = \frac{\lambda}{2\pi} \exp(\lambda x) \sin(2\pi y) \tag{19.51}$$

$$p(x, y) = \frac{1}{2}(1 - \exp(2\lambda x)) \tag{19.52}$$

$$\lambda = \frac{Re}{2} - \sqrt{\frac{Re^2}{4} + 4\pi^2} \tag{19.53}$$

Following step-55 of deal.II [1] we have to 'cheat' here since we are not solving the non-linear Navier-Stokes equations, but the linear Stokes system without convective term. Therefore, to recreate the exact same solution we move the convective term into the right-hand side.

The analytical solution is prescribed left and right, while free/no (??) slip is prescribed at top and bottom.

Velocity and pressure solution as implemented in step-55:

---

[1] https://www.dealii.org/current/doxygen/deal.II/step_55.html

```
const double pi2 = pi*pi;

u = −exp(x*(−sqrt(25.0 + 4*pi2) + 5.0))*cos(2*y*pi) + 1

v = (1.0L/2.0L)*(−sqrt(25.0 + 4*pi2) + 5.0)*
    exp(x*(−sqrt(25.0 + 4*pi2) + 5.0))*sin(2*y*pi)/pi

p = −1.0L/2.0L*exp(x*(−2*sqrt(25.0 + 4*pi2) + 10.0)) − 2.0*(−6538034.74494422
  + 0.0134758939981709*exp(4*sqrt(25.0 + 4*pi2)))/(−80.0*exp(3*sqrt(25.0 + 4*pi2))
  + 16.0*sqrt(25.0 + 4*pi2)*exp(3*sqrt(25.0 + 4*pi2)))
  − 1634508.68623606*exp(−3.0*sqrt(25.0 + 4*pi2))/(−10.0 + 2.0*sqrt(25.0 + 4*pi2))
  + (−0.00673794699908547*exp(sqrt(25.0 + 4*pi2))
  + 3269017.37247211*exp(−3*sqrt(25.0 + 4*pi2)))/(−8*sqrt(25.0 + 4*pi2) + 40.0)
  + 0.00336897349954273*exp(1.0*sqrt(25.0 + 4*pi2))/(−10.0 + 2.0*sqrt(25.0 + 4*pi2))
```

while the rhs of the PDE is given by

```
const double pi2 = pi * pi;

values[0] = −1.0L / 2.0L * (−2 * sqrt(25.0 + 4 * pi2) + 10.0) *
            exp(x*(−2*sqrt(25.0 + 4 * pi2) + 10.0)) −
            0.4 *pi2*exp(x * (−sqrt(25.0 + 4 * pi2) + 5.0)) * cos(2 * y * pi) +
            0.1 *pow(−sqrt(25.0 + 4 * pi2) + 5.0, 2) *
            exp(x*(−sqrt(25.0 + 4 * pi2) + 5.0)) * cos(2 * y * pi)

values[1] = 0.2 * pi*(−sqrt(25.0 + 4 * pi2) + 5.0) *
            exp(x*(−sqrt(25.0 + 4 * pi2) + 5.0)) * sin(2 * y * pi) −
            0.05 *pow(−sqrt(25.0 + 4 * pi2) + 5.0, 3) *
            exp(x*(−sqrt(25.0 + 4 * pi2) + 5.0)) * sin(2 * y * pi) / pi

values[2] = 0;
```
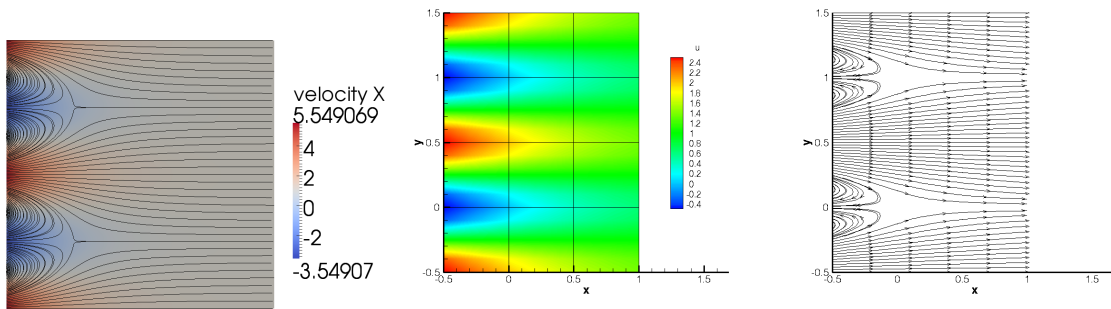

Left: solution from Step-55. Right: Solution obtained with NekTar++[2]

This benchmark is carried out in many CFD papers: [287, 125, 993], see also Section 7.4.3 of Hesthaven & Warburton [601].

Find analytical expression for pressure. Compute expression for rhs. Make stone

19.14   Manufactured solution in Lamichhane [784] (2017)

Taken from Lamichhane (2017) [784].

---
[2]http://doc.nektar.info/userguide/4.3.4/user-guidese45.html

$$u(x,y) = -2x^2y(2y-1)(x-1)^2(y-1) = a(x)b(y) \tag{19.54}$$
$$v(x,y) = 2xy^2(2x-1)(x-1)(y-1)^2 = c(x)d(y) \tag{19.55}$$
$$p(x,y) = x(1-x)(1-2y) \tag{19.56}$$

with

$$a(x) = -2x^2(x-1)^2 \tag{19.57}$$
$$b(y) = y(2y-1)(y-1) \tag{19.58}$$
$$c(x) = x(2x-1)(x-1) \tag{19.59}$$
$$d(y) = 2y^2(y-1)^2 \tag{19.60}$$

and

$$a'(x) = -4x(2x^2 - 3x + 1) \tag{19.61}$$
$$a''(x) = -4(6x^2 - 6x + 1) \tag{19.62}$$
$$b'(y) = 6y^2 - 6y + 1 \tag{19.63}$$
$$b''(y) = 12y - 6 \tag{19.64}$$
$$c'(x) = 6x^2 - 6x + 1 \tag{19.65}$$
$$c''(x) = 12x - 6 \tag{19.66}$$
$$d'(y) = 4y(2y^2 - 3y + 1) \tag{19.67}$$
$$d''(y) = 4(6y^2 - 6y + 1) \tag{19.68}$$

It then follows:

$$L_{xx} = \frac{\partial u}{\partial x} = a'(x)b(y) \tag{19.69}$$

$$L_{xy} = \frac{\partial v}{\partial x} = c'(x)d(y) \tag{19.70}$$

$$L_{yx} = \frac{\partial u}{\partial y} = a(x)b'(y) \tag{19.71}$$

$$L_{yy} = \frac{\partial v}{\partial y} = c(x)d'(y) \tag{19.72}$$

and

$$\dot{\varepsilon}_{xx} = L_{xx} = a'(x)b(y) \tag{19.73}$$
$$\dot{\varepsilon}_{yy} = L_{yy} = c(x)d'(y) \tag{19.74}$$
$$\dot{\varepsilon}_{xy} = \frac{1}{2}(L_{xy} + L_{yx}) \tag{19.75}$$
$$= \frac{1}{2}(a(x)b'(y) + c'(x)d(y)) \tag{19.76}$$

We easily verify that $\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} = 0$.

$$\frac{\partial p}{\partial x} = (1-2x)(1-2y) \tag{19.77}$$

$$\frac{\partial p}{\partial y} = -2x(1-x) \tag{19.78}$$

We will also need:

$$\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} = a''(x)b(y) \tag{19.79}$$

$$\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} = \frac{1}{2}(a(x)b''(y) + c'(x)d'(y)) \tag{19.80}$$

$$\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} = \frac{1}{2}(a'(x)b'(y) + c''(x)d(y)) \tag{19.81}$$

$$\frac{\partial \dot{\varepsilon}_{yy}}{\partial x} = c(x)d''(y) \tag{19.82}$$

Assuming $\eta = 1$ we arrive at:

$$b_x = \frac{\partial p}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} \tag{19.83}$$

$$= (1 - 2x)(1 - 2y) - 2a''(x)b(y) - (a(x)b''(y) + c'(x)d'(y)) \tag{19.84}$$

$$b_y = \frac{\partial p}{\partial y} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} \tag{19.85}$$

$$= -2x(1 - x) - (a'(x)b'(y) + c''(x)d(y)) - 2c(x)d''(y) \tag{19.86}$$

We have (thank you WolframAlpha[3]):

$$\int_0^1 \int_0^1 u^2 dxdy = \int_0^1 \int_0^1 (-2x^2 y(2y - 1)(x - 1)^2 (y - 1))^2 dxdy = \frac{1}{33075}$$

$$\int_0^1 \int_0^1 v^2 dxdy = \int_0^1 \int_0^1 (2xy^2(2x - 1)(x - 1)(y - 1)^2)^2 dxdy = \frac{1}{33075}$$

so that

$$v_{rms} = \sqrt{\frac{1}{L_x L_y} \frac{2}{33075}} \simeq 0.00777615791$$

Finally we can verify that the pressure has zero average:

$$\int_0^1 \int_0^1 p(x, y) dxdy = \int_0^1 \int_0^1 x(1 - x)(1 - 2y) dxdy = \int_0^1 x(1 - x)dx \int_0^1 (1 - 2y)dy = 0$$

19.15   Manufactured solution in Mu and Ye [965]

We first consider the Stokes equations with homogeneous boundary condition defined on the unit square. The exact solutions are given by

$$
\begin{aligned}
u(x, y) &= 2\pi \sin^2(\pi x) \cos(\pi y) \sin(\pi y) \\
v(x, y) &= -2\pi \sin(\pi x) \cos(\pi x) \sin^2(\pi y) \\
p(x, y) &= \cos(\pi x) \cos(\pi y)
\end{aligned} \tag{19.87}
$$

Then, since $\cos' = -\sin$:

$$
\begin{aligned}
\partial_x p &= -\pi \sin(\pi x) \cos(\pi y) \\
\partial_y p &= -\pi \cos(\pi x) \sin(\pi y)
\end{aligned}
$$

---

[3]

and

$$\begin{aligned}
\partial_x u &= 4\pi^2 \sin(\pi x) \cos(\pi x) \cos(\pi y) \sin(\pi y) \\
\partial_y u &= 2\pi^2 \sin^2(\pi x)[\cos^2(\pi y) - \sin^2(\pi y)] \\
\partial_x v &= -2\pi^2[\cos^2(\pi x) - \sin^2(\pi x)] \sin^2(\pi y) \\
\partial_y v &= -4\pi^2 \sin(\pi x) \cos(\pi x) \sin(\pi y) \cos(\pi y)
\end{aligned}$$

with as expected

$$\partial_x u + \partial_y v = 4\pi^2 \sin(\pi x) \cos(\pi x) \cos(\pi y) \sin(\pi y) - 4\pi^2 \sin(\pi x) \cos(\pi x) \sin(\pi y) \cos(\pi y) = 0$$

The strain rate components are then

$$\begin{aligned}
\dot{\varepsilon}_{xx} &= \partial_x u \\
\dot{\varepsilon}_{yy} &= \partial_y v \\
\dot{\varepsilon}_{xy} &= \frac{1}{2}(\partial_x v + \partial_y u) \\
&= \frac{1}{2}\left(-2\pi^2[\cos^2(\pi x) - \sin^2(\pi x)] \sin^2(\pi y) + 2\pi^2 \sin^2(\pi x)[\cos^2(\pi y) - \sin^2(\pi y)]\right) \\
&= \pi^2[\sin^2(\pi x) - \cos^2(\pi x)] \sin^2(\pi y) + \pi^2 \sin^2(\pi x)[\cos^2(\pi y) - \sin^2(\pi y)] \\
&= \pi^2\left(\sin^2(\pi x) \cos^2(\pi y) - \cos^2(\pi x) \sin^2(\pi y)\right)
\end{aligned}$$

We will also need

$$\begin{aligned}
\partial_x \dot{\varepsilon}_{xx} &= 4\pi^3[\cos^2(\pi x) - \sin^2(\pi x)] \cos(\pi y) \sin(\pi y) \\
\partial_x \dot{\varepsilon}_{xy} &= \pi^3\left(2\cos(\pi x) \sin(\pi x) \cos^2(\pi y) + 2\sin(\pi x) \cos(\pi x) \sin^2(\pi y)\right) \\
&= 2\pi^3 \sin(\pi x) \cos(\pi x) \\
\partial_y \dot{\varepsilon}_{xy} &= \pi^3\left(-2\sin^2(\pi x) \sin(\pi y) \cos(\pi y) - 2\cos^2(\pi x) \sin(\pi y) \cos(\pi y)\right) \\
&= -2\pi^3 \sin(\pi y) \cos(\pi y) \\
\partial_y \dot{\varepsilon}_{yy} &= -4\pi^3 \sin(\pi x) \cos(\pi x)[\cos^2(\pi y) - \sin^2(\pi y)]
\end{aligned}$$

Finally we need to compute the root mean square velocity (integrals were obtained with Wolframalpha):

$$\iint u^2 dx dy = 4\pi^2 \underbrace{\int_0^1 \sin^4(\pi x) dx}_{3/8} \cdot \underbrace{\int_0^1 \cos^2(\pi y) \sin^2(\pi y) dy}_{1/8}$$

$$\iint v^2 dx dy = 4\pi^2 \underbrace{\int_0^1 \sin^2(\pi x) \cos^2(\pi x) dx}_{1/8} \cdot \underbrace{\int_0^1 \sin^4(\pi y) dy}_{3/8}$$

so that

$$\nu_{rms} = \sqrt{\frac{1}{L_x L_y} \iint (u^2 + v^2) dx dy} = \sqrt{\frac{3}{8}\frac{1}{8} + \frac{1}{8}\frac{3}{8}} = \sqrt{\frac{3}{32}} \simeq 0.05412658773$$

Python script:

19.16 Manufactured solution in Boffi, Cavallini, Gardini, and Gastaldi [116] (2012)

This manufactured solution originates in Boffi, Cavallini, Gardini, and Gastaldi [116] (2012). The velocity field turns out to be identical to the Donea and Huerta [363] manufactured solution. It is based on the stream function

$$\Psi(x, y) = x^2(x-1)^2 y^2(y-1)^2 = f(x)g(y)$$

defined on the unit square. We have

$$
\begin{aligned}
f'(x) &= 2x(x-1)^2 + 2x^2(x-1) \\
&= 2(x-1)[x(x-1) + x^2] \\
&= 2(x-1)(2x^2 - x) \\
&= 2x(x-1)(2x-1) \\
f''(x) &= 2[(x-1)(2x-1) + x(2x-1) + 2x(x-1)] \\
&= 2(2x^2 - x - 2x + 1 + 2x^2 - x + 2x^2 - 2x) \\
&= 2(6x^2 - 6x + 1) \\
g'(y) &= 2y(y-1)^2 + 2y^2(y-1) \\
&= 2(y-1)[y(y-1) + y^2] \\
&= 2(y-1)(2y^2 - y) \\
&= 2y(y-1)(2y-1) \\
g''(y) &= 2[(y-1)(2y-1) + y(2y-1) + 2y(y-1)] \\
&= 2[2y^2 - 3y + 1 + 2y^2 - y + 2y^2 - 2y] \\
&= 2(6y^2 - 6y + 1)
\end{aligned}
\tag{19.88}
$$

The manufactured 'smooth pressure' solution is then

$$
\begin{aligned}
u(x, y) &= \partial_y \psi = f(x)g'(y) = x^2(x-1)^2 2y(y-1)(2y-1) \\
v(x, y) &= -\partial_x \psi = -f'(x)g(y) = -2x(x-1)(2x-1)y^2(y-1)^2 \\
p(x, y) &= \frac{1}{2}x^2 - \frac{1}{6}
\end{aligned}
$$

$$
\begin{aligned}
\partial_x u(x, y) &= f'(x)g'(y) \\
\partial_y u(x, y) &= f(x)g''(y) \\
\partial_x v(x, y) &= -f''(x)g(y) \\
\partial_y v(x, y) &= -f'(x)g'(y)
\end{aligned}
$$

$$
\begin{aligned}
\dot{\varepsilon}_{xx}(x, y) &= f'(x)g'(y) \\
\dot{\varepsilon}_{yy}(x, y) &= -f'(x)g'(y) \\
\dot{\varepsilon}_{xy}(x, y) &= \frac{1}{2}(f(x)g''(y) - f''(x)g(y))
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial p}{\partial x}(x, y) &= x \\
\frac{\partial p}{\partial y}(x, y) &= 0
\end{aligned}
$$

$$\partial_x \dot{\varepsilon}_{xx}(x, y) = f''(x)g'(y)$$

$$\partial_x \dot{\varepsilon}_{xy}(x, y) = \frac{1}{2}\left(f'(x)g''(y) - f'''(x)g(y)\right)$$

$$\partial_y \dot{\varepsilon}_{xy}(x, y) = \frac{1}{2}\left(f(x)g'''(y) - f''(x)g'(y)\right)$$

$$\partial_y \dot{\varepsilon}_{yy}(x, y) = -f'(x)g''(y)$$

$$\boldsymbol{\nu}_{rms} = \sqrt{\frac{1}{L_x L_y} \iint (u^2 + v^2)dxdy}$$

$$= \sqrt{\frac{1}{L_x L_y} \iint (f^2 g'^2 + f'^2 g^2)dxdy}$$

$$= \sqrt{\left(\underbrace{\int_0^1 f^2 dx}_{1/630} \underbrace{\int_0^1 g'^2 dy}_{2/105} + \underbrace{\int_0^1 f'^2 dx}_{2/105} \underbrace{\int_0^1 g^2 dy}_{1/630}\right) dxdy}$$

$$1/33075 * 2 = \frac{1}{105}\sqrt{\frac{2}{3}}$$

$$\simeq 0.0077761579135973907879278859 51... \tag{19.89}$$

The authors also define a non-smooth pressure case:

$$p(x, y) = \begin{cases} y(1-y)\exp(x - 1/2)^2 + 1/2 & \text{for} \quad x \geq 1/2 \\ y(1-y)\exp(x - 1/2)^2 - 1/2 & \text{for} \quad x < 1/2 \end{cases}$$

with

$$\frac{\partial p}{\partial x}(x, y) = 2(x - 1/2)y(1 - y)\exp(x - 1/2)^2$$

$$\frac{\partial p}{\partial y}(x, y) = (1 - 2y)\exp(x - 1/2)^2$$

However it is not clear to me how to implement this since only the gradient of the pressure appears in the equations.

19.17   Manufactured solution in John [688] (book)

This manufactured solution originates in appendix D.1 of John [688] (book).
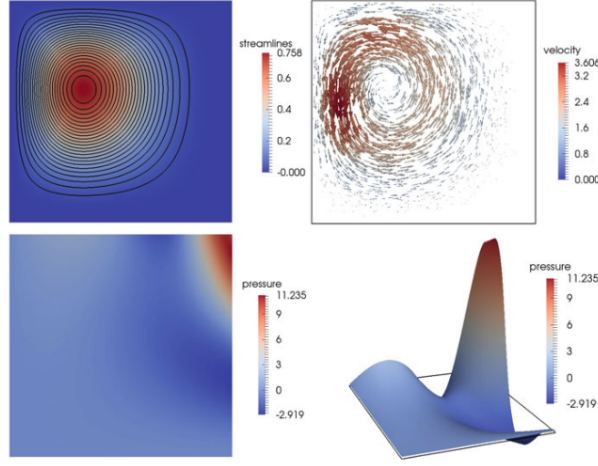
Fig. D.1 Example D.3. Stream function (*top left*) velocity (*top right*) and pressure (*bottom*). These plots are based on results obtained with numerical simulations

Taken from [688].

The stream function is given by

$$\Phi(x, y) = 1000x^2(1-x)^4 y^3(1-y)^2 = 1000f(x)g(y)$$

with

$$
\begin{aligned}
f(x) &= x^2(1-x)^4 \\
f'(x) &= 2x(1-x)^4 - 4x^2(1-x)^3 \\
&= [2x(1-x) - 4x^2](1-x)^3 \\
&= (2x - 6x^2)(1-x)^3 \\
&= 2x(1-3x)(1-x)^3 \\
f''(x) &= 2(1-3x)(1-x)^3 - 6x(1-x)^3 - 6x(1-3x)(1-x)^2 \\
&= 2[(1-3x)(1-x) - 3x(1-x) - 3x(1-3x)](1-x)^2 \\
&= 2[1 - 4x + 3x^2 - 3x + 3x^2 - 3x + 9x^2](1-x)^2 \\
&= 2(1 - 10x + 15x^2)(1-x)^2 \\
f'''(x) &= 2[(-10 + 30x)(1-x)^2 - 2(1 - 10x + 15x^2)(1-x)] \\
&= 2[-10 + 40x - 30x^2 - 2 + 20x - 30x^2](1-x) \\
&= 2(-12 + 60x - 60x^2)(1-x) \\
&= 24(-1 + 5x - 5x^2)(1-x) \\
&= 24(-1 + 5x - 5x^2 + x - 5x^2 + 5x^3) \\
&= 24(-1 + 6x - 10x^2 + 5x^3)
\end{aligned}
$$

$$
\begin{aligned}
g(y) &= y^3(1-y)^2 \\
g'(y) &= 3y^2(1-y)^2 - 2y^3(1-y) \\
&= [3y^2(1-y) - 2y^3](1-y) \\
&= y^2(3 - 3y - 2y)(1-y) \\
&= y^2(3 - 5y)(1-y) \\
g''(y) &= 2y(3 - 5y)(1-y) - 5y^2(1-y) - y^2(3 - 5y) \\
&= 2y(3 - 8y + 5y^2) - 5y^2 + 5y^3 - 3y^2 + 5y^3 \\
&= 6y - 16y^2 + 10y^3 - 8y^2 + 10y^3 \\
&= 2y(3 - 12y + 10y^2) \\
g'''(y) &= 6 - 48y + 60y^2
\end{aligned}
$$

984

$$
\begin{aligned}
u(x,y) &= \partial_y \Phi = 1000 f(x) g'(y) = 1000 x^2 (1-x)^4 y^2 (3-5y)(1-y) & (19.90)\\
v(x,y) &= -\partial_x \Phi = -1000 f'(x) g(y) = -10002 x(1-3x)(1-x)^3 y^3 (1-y)^2 & (19.91)\\
p(x,y) &= \pi^2 [xy^3 \cos(2\pi x^2 y) - x^2 y \sin(2\pi xy)] + 1/8 & (19.92)
\end{aligned}
$$
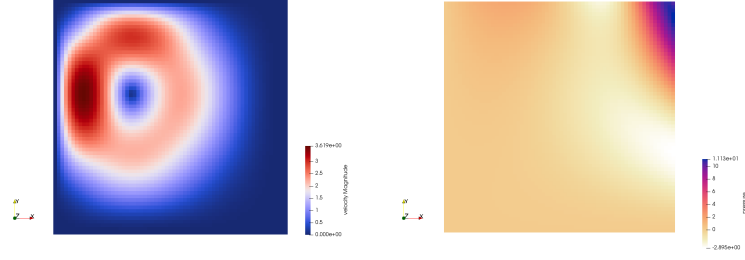
$$
\begin{aligned}
\partial_x u(x,y) &= 1000 f'(x) g'(x)\\
\partial_y u(x,y) &= 1000 f(x) g''(y)\\
\partial_x v(x,y) &= -1000 f''(x) g(y)\\
\partial_y v(x,y) &= -1000 f'(x) g'(y)\\
\dot{\varepsilon}_{xx}(x,y) &= 1000 f' g'\\
\dot{\varepsilon}_{yy}(x,y) &= -1000 f' g'\\
\dot{\varepsilon}_{xy}(x,y) &= 500(fg'' - f''g)\\
\partial_x \dot{\varepsilon}_{xx}(x,y) &= 1000 f'' g'\\
\partial_x \dot{\varepsilon}_{xy}(x,y) &= 500(f'g'' - f'''g)\\
\partial_y \dot{\varepsilon}_{xy}(x,y) &= 500(fg''' - f''g')\\
\partial_y \dot{\varepsilon}_{yy}(x,y) &= -1000 f' g''
\end{aligned}
$$

Of course we have $\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} = 0$.

$$
\begin{aligned}
\frac{\partial p}{\partial x}(x,y) &= \pi^2[y^3 \cos(2\pi x^2 y) - 4\pi x^2 y^4 \sin(2\pi x^2 y) - 2xy\sin(2\pi xy) - 2\pi x^2 y^2 \cos(2\pi xy)]\\
\frac{\partial p}{\partial y}(x,y) &= \pi^2[3xy^2 \cos(2\pi x^2 y) - 2\pi x^3 y^3 \sin(2\pi x^2 y) - x^2 \sin(2\pi xy) - 2\pi x^3 y \cos(2\pi xy)]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{v}_{rms} &= \sqrt{\frac{1}{L_x L_y} \iint (u^2 + v^2) dx dy}\\[2mm]
&= \sqrt{\int_0^1 \int_0^1 u^2 \, dx dy + \int_0^1 \int_0^1 v^2 \, dx dy}\\[2mm]
&= 1000 \sqrt{\int_0^1 \int_0^1 (fg')^2 \, dx dy + \int_0^1 \int_0^1 (-f'g)^2 \, dx dy}\\[2mm]
&= 1000 \sqrt{\underbrace{\int_0^1 f^2 dx}_{\frac{1}{6435}} \underbrace{\int_0^1 g'^2 dy}_{\frac{2}{315}} + \underbrace{\int_0^1 f'^2 dx}_{\frac{2}{693}} \underbrace{\int_0^1 g^2 dy}_{\frac{1}{2310}}}\\[2mm]
&= 1000 \sqrt{\frac{1}{5\cdot 9\cdot 13\cdot 11}\frac{2}{5\cdot 9\cdot 7} + \frac{1}{9\cdot 11\cdot 7}\frac{1}{5\cdot 11\cdot 3\cdot 7}}\\[2mm]
&\simeq 1.4953325891041323968540981 & (19.93)
\end{aligned}
$$

19.18 Manufactured solution in John, Kaiser, and Novo [690]

This benchmark is identical to the one presented in Section 19.17 but it is not isoviscous. There are three different viscosity fields:

$$\eta_1(x,y) = \eta_{min} + (\eta_{max} - \eta_{min})x^2(1-x)y^2(1-y)\frac{721}{16} \tag{19.94}$$

$$\eta_2(x,y) = \eta_{min} + (\eta_{max} - \eta_{min})\exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}] \tag{19.95}$$

$$\eta_3(x,y) = \eta_{min} + (\eta_{max} - \eta_{min})\left[1 - \exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}]\right] \tag{19.96}$$
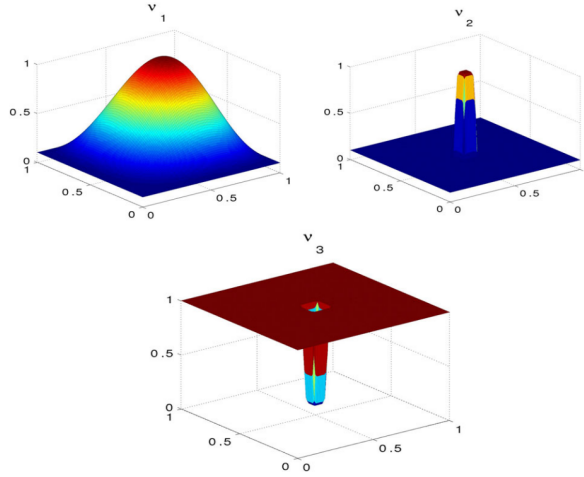
Taken from John, Kaiser, and Novo [690] (2016).

We have

$$b_x = \frac{\partial p}{\partial x} - 2\eta\frac{\partial\dot{\varepsilon}_{xx}}{\partial x} - 2\eta\frac{\partial\dot{\varepsilon}_{xy}}{\partial y} - 2\frac{\partial\eta}{\partial x}\dot{\varepsilon}_{xx} - 2\frac{\partial\eta}{\partial y}\dot{\varepsilon}_{xy} \tag{19.97}$$

$$b_y = \frac{\partial p}{\partial y} - 2\eta\frac{\partial\dot{\varepsilon}_{xy}}{\partial x} - 2\eta\frac{\partial\dot{\varepsilon}_{yy}}{\partial y} - 2\frac{\partial\eta}{\partial x}\dot{\varepsilon}_{xy} - 2\frac{\partial\eta}{\partial y}\dot{\varepsilon}_{yy} \tag{19.98}$$

The three first terms on each line have already been obtained so we must focus on the last two:

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_1}{\partial x} = \frac{721}{16}[2x(1-x) - x^2]y^2(1-y) \tag{19.99}$$

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_1}{\partial y} = \frac{721}{16}x^2(1-x)[2y(1-y) - y^2] \tag{19.100}$$

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_2}{\partial x} = -10^{14}(x-0.5)^9\exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}] \tag{19.101}$$

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_2}{\partial y} = -10^{14}(y-0.5)^9\exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}] \tag{19.102}$$

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_3}{\partial x} = 10^{14}(x-0.5)^9\exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}] \tag{19.103}$$

$$\frac{1}{(\eta_{max} - \eta_{min})}\frac{\partial\eta_3}{\partial y} = 10^{14}(y-0.5)^9\exp[-10^{13}(x-0.5)^{10} + (y-0.5)^{10}] \tag{19.104}$$

with

$$u(x,y) = 1000x^2(1-x)^4y^2(3-5y)(1-y) \tag{19.105}$$

$$v(x,y) = -10002x(1-3x)(1-x)^3y^3(1-y)^2 \tag{19.106}$$

$$p(x,y) = \pi^2[xy^3\cos(2\pi x^2 y) - x^2 y\sin(2\pi xy)] + 1/8 \tag{19.107}$$

Note that between John's book and the papers [690] and [691] there seems to be a small difference: xy^2 vs. xy^3. Also the papers have -1/8 while it should be +1/8 (thank you Wolfram Alpha).

19.19   Manufactured solution in John [687] (1998) on a disc

Let Ω be the disc with the center (0,0(and the radius 1 which has a crack along the $x-$axis between the points (0,0) and (1,0).

$$\vec{v}(x,y) = \frac{3}{2}\sqrt{r}\left(\cos\frac{\theta}{2} - \cos\frac{3\theta}{2}, 3\sin\frac{\theta}{2} - \sin\frac{3\theta}{2}\right)$$

$$p(x,y) = -\frac{6}{\sqrt{r}}\cos\frac{\theta}{2}$$

Viscosity is 1. Note that the solution has a singularity in the origin.
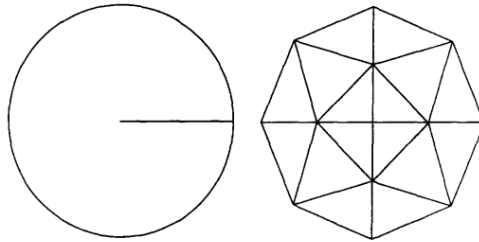
Fig. 4. Domain and coarsest grid (level 0) for Example 11.
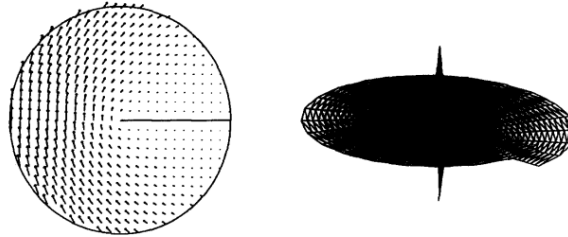
Fig. 5. Solution (velocity and pressure) of Example 11.

Taken from John [687] (1998).

19.20   Manufactured solution in Böhm [117] (2023)

The author does not specify the size of the domain nor the boundary conditions.

$$\vec{v}(x,y) = \begin{pmatrix} y + 2\sin(\pi(x+y)) + 4 \\ -x - 2\sin(\pi(x+y)) + 3 \end{pmatrix} \qquad p(x,y) = 2x - y + 1$$

Since the velocity is not zero on the sides of the unit square and never will on any straight horizontal or vertical line, we will assume that the velocity is prescribed on the sides. Note also that the velocity is not normalised to zero:

$$\iint p(x,y)dxdy = 0.5$$

The viscosity function is non-constant but smooth:

$$\eta(x,y) = \exp(x)\sin\left(\pi\frac{x+1}{2}\right)\sin\left(\pi\frac{y+1}{2}\right) + 1$$

The authors run experiments with $P_2 \times P_1$ element pair and recover expected error convergence rates.

19.21 Driven cavity with analytical solution

This comes from Elman *et al.* [394](section 3.1.4)[4]. The velocity is prescribed to be

$$\vec{v} = (2y(1-x^2); -2x(1-y^2))$$

on the domain $\Omega = [-1 : 1] \times [-1 : 1]$. The strainrate tensor is given by:

$$\dot{\varepsilon} = \begin{pmatrix} -4xy & -x^2 + y^2 \\ -x^2 + y^2 & 4xy \end{pmatrix}$$

The Stokes equation is then:

$$-\frac{\partial p}{\partial x} + 2\eta(-4y + 2y) = \rho g_x \tag{19.108}$$

$$-\frac{\partial p}{\partial y} + 2\eta(-2x + 4x) = \rho g_y \tag{19.109}$$

where we assume the viscosity $\eta = 1$ to be constant in space. Assuming $g_x = 0$, the first equation is

$$\frac{\partial p}{\partial x} = -4y$$

i.e.

$$p(x, y) = -4yx + f(y)$$

Inserting this in the second equation:

$$4x - f'(y) + 4x = \rho g_y$$

or,

$$-f'(y) + 8x = \rho g_y$$

Assuming $g_y = -1$, we get $\rho = -8x$ and then $f'(y) = 0$ so $f(y) = C$ where C is a constant. Finally the pressure is given by:

$$p(x, y) = -4yx + C$$

We add the following requirement: $\int_\Omega p(x, y)d\Omega = 0$ so that $C = 0$.

---
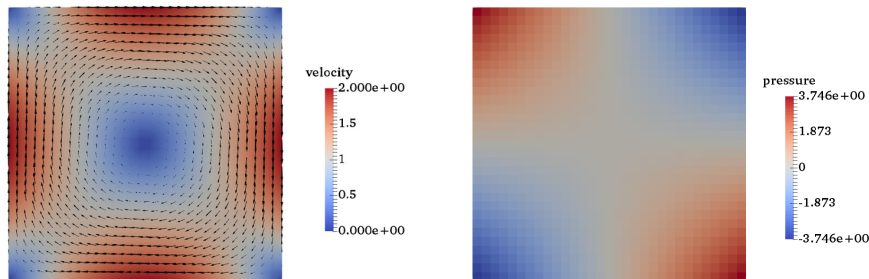
[4]actually, not?

$$
\begin{aligned}
\mathbf{v}_{rms}^2 &= \frac{1}{\Omega} \int_\Omega (u^2 + v^2) d\Omega \\
&= \frac{1}{4} \int_{-1}^{+1} \int_{-1}^{+1} (u^2 + v^2) dx dy \\
&= \frac{1}{4} \int_{-1}^{+1} \int_{-1}^{+1} [4y^2(1-x^2)^2 + 4x^2(1-y^2)^2] dx dy \\
&= \int_{-1}^{+1} \int_{-1}^{+1} [y^2(1-x^2)^2 + x^2(1-y^2)^2] dx dy \\
&= \int_{-1}^{+1} \int_{-1}^{+1} [y^2(1-2x+x^2) + x^2(1-2y+y^2)] dx dy \\
&= \int_{-1}^{+1} \int_{-1}^{+1} y^2 dx dy + \int_{-1}^{+1} \int_{-1}^{+1} 2x^2 y^2 dx dy + \int_{-1}^{+1} \int_{-1}^{+1} x^2 dx dy \\
&= 2\frac{2}{3} + 2\frac{2}{3}\frac{2}{3} + 2\frac{2}{3} \\
&= \frac{32}{9} \\
&= 3.5555 \tag{19.110}
\end{aligned}
$$

We can reformulate the benchmark in the unit square $\Omega = [0:1] \times [0:1]$.

$$
\begin{aligned}
u(x,y) &= (2y-1)x(1-x) \\
v(x,y) &= -(2x-1)y(1-y)
\end{aligned}
$$

Then

$$
\begin{aligned}
\dot{\varepsilon}_{xx}(x,y) &= (2y-1)(1-2x) \\
\dot{\varepsilon}_{xy}(x,y) &= x(1-x) - y(1-y) \\
\dot{\varepsilon}_{yy}(x,y) &= -(2x-1)(1-2y)
\end{aligned}
$$

We of course recover

$$
\dot{\varepsilon}_{xx}(x,y) + \dot{\varepsilon}_{yy}(x,y) = 0
$$

The Stokes equation is then:

$$
\begin{aligned}
-\frac{\partial p}{\partial x} + 2\eta(-2(2y-1) - (1-2y)) &= \rho g_x \\
-\frac{\partial p}{\partial y} + 2\eta((1-2x) + 2(2x-1)) &= \rho g_y
\end{aligned}
$$

or

$$
\begin{aligned}
-\frac{\partial p}{\partial x} + 2\eta(1-2y) &= \rho g_x \\
-\frac{\partial p}{\partial y} - 2\eta(1-2x) &= \rho g_y
\end{aligned}
$$

where we assume the viscosity $\eta = 1$ to be constant in space. Assuming $g_x = 0$, the first equation is

$$
\frac{\partial p}{\partial x} = 2(1-2y)
$$

989

i.e.
$$p(x, y) = 2x(1 - 2y) + f(y)$$
Inserting this in the second equation:
$$4x - f'(y) - 2(1 - 2x) = \rho g_y$$
Assuming $g_y = -1$, then
$$8x - 2 - f'(y) = \rho$$
we set $\rho(x, y) = 8x - 2$ and then $f'(y) = 0$ so $f(y) = C$ where C is a constant. Finally the pressure is given by:
$$p(x, y) = 2x(1 - 2y) + C$$
We add the following requirement: $\int_\Omega p(x, y)d\Omega = 0$
$$\int_0^1 \int_0^1 p(x, y) = 0 \Rightarrow \int_0^1 \int_0^1 [2x(1 - 2y) + C] = 0$$
so that $C = 0$.

We also find $\mathbf{v}_{rms} = \frac{1}{\sqrt{45}} \simeq 0.1490711985$

19.22 Annulus with kinematical b.c. - pure rotation

The domain is a hollow cylinder or inner radius R_i and outside radius $R_o = 1$. Boundary conditions are prescribed both on the inside and the outside with $\vec{v} = (u, v) = (-y, x)$, or in polar coordinates $\vec{v} = r\vec{e}_\theta$.

The gravity is radial and is set to
$$g_x = -x/r \qquad g_z = -y/r$$
where $r = \sqrt{x^2 + z^2}$, which in polar coordinates is $\vec{g} = -\vec{e}_r$. The viscosity is also set to 1, and the density is given by
$$\rho(r) = r^n$$
where n is a positive or nul integer. The pressure is set to zero at the outer boundary.

The gradient operator in polar coordinates writes:
$$\vec{\nabla} = \frac{\partial}{\partial r}\vec{e}_r + \frac{1}{r}\frac{\partial}{\partial \theta}\vec{e}_\theta$$
and the Laplacian operator:
$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}$$
Note that in our case we need to take the Laplacian of a vector, and unfortunately the Laplacian of a vector is not the Laplacian of the vector's coordinates in polar coordinates (unlike cartesian coordinates). The Laplacian of a vector is given by[5]

$$\nabla^2 \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla \times (\nabla \times \vec{A}) = \begin{pmatrix} \frac{\partial^2 A_r}{\partial r^2} + \frac{1}{r}\frac{\partial A_r}{\partial r} - \frac{1}{r^2}A_r + \frac{1}{r^2}\frac{\partial^2 A_r}{\partial \theta^2} - \frac{2}{r^2}\frac{\partial A_\theta}{\partial \theta} \\ \\ \frac{\partial^2 A_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial A_\theta}{\partial r} - \frac{1}{r^2}A_\theta + \frac{1}{r^2}\frac{\partial^2 A_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial A_r}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \Delta A_r \\ \\ \Delta A_\theta \end{pmatrix}$$

---

[5]https://en.wikipedia.org/wiki/Vector_Laplacian

The Stokes equation writes:
$$-\vec{\nabla}p + \eta\Delta\vec{v} + \rho\vec{g} = \vec{0}$$

The velocity solution is expected to be $\vec{v} = r\vec{e}_\theta$. The Stokes equation in polar coordinates then writes:

$$-\frac{\partial p}{\partial r} + \Delta v_r + \rho(r)(-1) = 0$$

$$-\frac{1}{r}\frac{\partial p}{\partial \theta} + \Delta v_\theta = 0$$

Since $\Delta v_\theta = 0$, then $\frac{\partial p}{\partial \theta} = 0$ and then the pressure is independent of θ, which is what we expect since the density distribution is radial. We then focus on the first equation, and since $v_r = 0$, we then obtain:

$$\frac{\partial p}{\partial r} = -\rho(r)$$

- If $\rho(r) = 1$, then
$$\frac{\partial p}{\partial r} = -1$$

  yields $p(r) = -r + C$ where C is a constant determined by means of b.c. $(p(r = 1) = 0)$ so finally

$$\boxed{p(r) = 1 - r}$$

- If $\rho(r) = r$, then
$$\frac{\partial p}{\partial r} = -r$$

  so that $p(r) = -\frac{1}{2}r^2 + C$ and likewise

$$\boxed{p(r) = \frac{1}{2}(1 - r^2)}$$

In general, by taking $\rho(r) = r^n$ with $n = 0, 1, ...$ one arrives to a pressure field given by

$$\boxed{p(r) = \frac{1}{n+1}(1 - r^{n+1})}$$
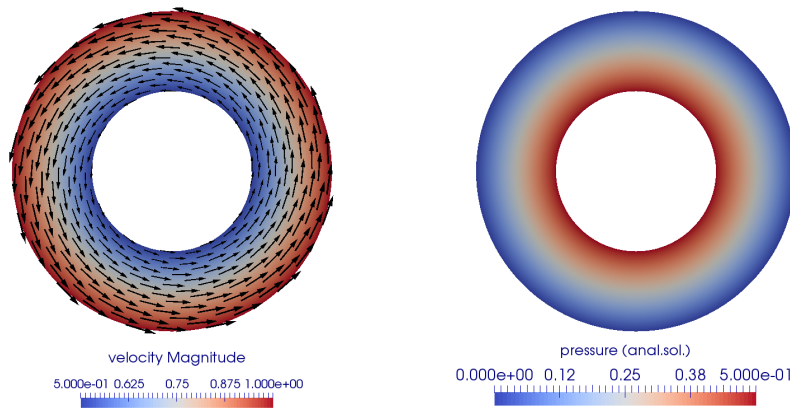
This benchmark is of course very simple and the fact that the solution is independent of θ renders it not so useful. It has succesfully been implemented in ELEFANT .

19.23 Viscous beam under extension

The domain is a Cartesian box of size $L_x \times L_y$. Velocity $-u_0$ is applied on the left boundary and velocity $+u_0$ is applied on the right boundary. Bottom and top boundaries are left free. If no vertical velocity is prescribed anywhere there is an obvious nullspace in the solution which is problematic (numerically of course, but also because the solution is then not unique). One might want to set $v = 0$ at $y = L_y/2$ on each side for example. The solution to this problem (incompressible Stokes equations) is given by

$$u(x, y) = 2u_0(x/L_x - 1/2) \tag{19.111}$$
$$v(x, y) = -2u_0 L_y/L_x(y/L_y - 1/2) \tag{19.112}$$

in the absence of gravity. The strain rate tensor is then:

$$\dot{\boldsymbol{\varepsilon}} = \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} \\ \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{yy} \end{pmatrix} = \begin{pmatrix} 2u_0/Lx & 0 \\ 0 & -2u_0/L_x \end{pmatrix}$$

and we see that the flow is indeed incompressible as the trace of the strain rate tensor is zero.

The momentum equation is

$$-\vec{\nabla}p + \vec{\nabla} \cdot (2\eta\dot{\boldsymbol{\varepsilon}}) = \rho\vec{g}$$

where the viscosity η is constant in space. If gravity is set to zero, we obtain:

$$-\frac{\partial p}{\partial x} = 0 \tag{19.113}$$
$$-\frac{\partial p}{\partial y} = 0 \tag{19.114}$$

since the strain rate is constant in space and the divergence operator applied to it returns the zero tensor. We there fore can conclude that pressure should be constant.

Since the top and bottom boundaries are free, we have $\boldsymbol{\sigma} \cdot \vec{n} = \vec{0}$ on these. The stress tensor is given by $\boldsymbol{\sigma} = -\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}$ and the normal on the top is $\vec{n} = (0, +1)$ so that on the top boundary we have

$$-p + 2\eta\dot{\epsilon}_{yy} = 0$$

or,

$$p = 2\eta\dot{\epsilon}_{yy}$$

Note that using the bottom boundary with $\vec{n} = (0, -1)$ yields the same result.

19.24 Simple shear heating

The domain is a $L_x \times L_y$ Cartesian box. The velocity field $\vec{v} = (L_y - y)y\vec{e}_x$ is prescribed on all boundaries, or simply prescribed everywhere in the domain. Temperature is set to $T = 0$ everywhere in the domain at $t = 0$.

As we have seen in Section 3.6, the shear heating, Φ is expressed as:

$$\Phi = 2\eta\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) : \dot{\boldsymbol{\varepsilon}}^d(\vec{v}) \tag{19.115}$$

We have

$$\dot{\varepsilon}_{xx}(\vec{v}) = \partial_x u = 0$$
$$\dot{\varepsilon}_{yy}(\vec{v}) = \partial_y v = 0$$
$$\dot{\varepsilon}_{xy}(\vec{v}) = \frac{1}{2}(\partial_x v + \partial_y u) = \frac{1}{2}(L_y - 2y) \tag{19.116}$$

We see that the flow is incompressible ($\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} = 0$) so that $\dot{\boldsymbol{\varepsilon}}^d(\vec{v}) = \dot{\boldsymbol{\varepsilon}}(\vec{v})$ and

$$\Phi(x,y) = 2\eta[\dot{\varepsilon}_{xx}(\vec{v})^2 + \dot{\varepsilon}_{yy}(\vec{v})^2 + 2\dot{\varepsilon}_{xy}(\vec{v})^2] = 2\eta 2\frac{1}{4}(L_y - 2y)^2 = \eta(L_y - 2y)^2$$

We set $\eta = 1$ and $L_y = 1$ (L_x is actually irrelevant) so that $\Phi(x,y) = (1-2y)^2$.

The energy equation is given by

$$\rho C_p \left(\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla} T \right) = k\Delta T + \Phi, \tag{19.117}$$

assuming that there is no conduction (i.e. $k=0$) and since $\vec{\nabla}T \propto \vec{e}_y$ while $\vec{v} \propto \vec{e}_x$ then it simplifies to

$$\frac{\partial T}{\partial t} = \Phi$$

where we have taken $\rho = 1$ and $C_p = 1$ for convenience. Since $T(t=0) = 0$, we then have

$$T(t) = \Phi\, t$$

19.25 2d solution with nontrivial interface jump

This benchmark is featured in Sevilla and Duretz [1228] (2023) but orginates in Wang and Khoo [1435] (2013). The domain is $\Omega = [0,2] \times [-0.5, 1.5]$ and the viscosity is given by $\eta(x,y) = \eta_1$ if $y \le 0.5$ and $\eta(x,y) = \eta_2$ if $y > 0.5$. The analytical solution is given by

$$\vec{v} = (1 - \exp(\lambda)\sin(2\pi y), 0) \qquad p = \frac{1}{2}\exp(2\lambda x) \qquad \text{with} \qquad \lambda = \frac{1}{2\eta} - \sqrt{\frac{1}{4\eta^2} + 4\pi^2} \tag{19.118}$$

The pressure and the gradient tensor exhibit a discontinuity across the interface Dirichlet boundary conditions are imposed in the whole boundary.
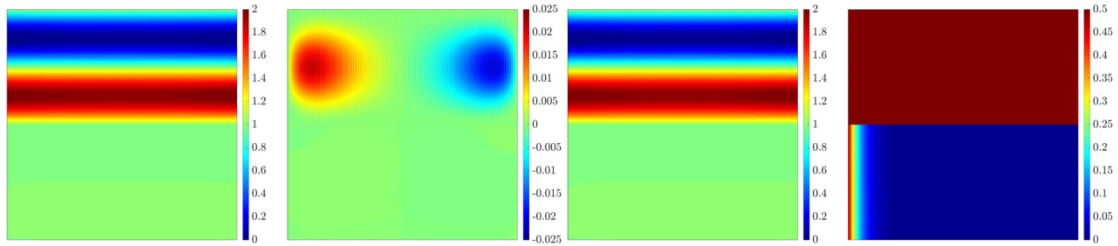
Taken from [1228]. From left to right: u, v, $|\vec{v}|$, p, for $\eta_1 = 1$ and $\eta = 10^{-4}$.

19.26 3d solution with nontrivial interface jump

This benchmark is taken from Sevilla and Duretz [1228] but is taken from Kirchhart, Gross, and Reusken [751] (2016). It is a three dimensional problem in $[-1,1]^3$ and the viscosity is given by $\eta(x,y) = \eta_1$ if $r \le r_I$ and $\eta(x,y) = \eta_2$ if $r > r_I$ where $r = |\vec{v}|_2$ and $r_I = 2/3$. The analytical solution is given by

$$\vec{v} = \alpha(r)\exp(-r^2)(-y, x, 0) \qquad p = x^3 + \lambda(r) \tag{19.119}$$

where

$$\alpha(r) = 1/\eta_1 \qquad\qquad\qquad if\ r \leq r_I \qquad (19.120)$$
$$= 1/\eta_2 + (1/\eta_1 - 1/\eta_2)\exp(r^2 - r_I^2) \qquad if\ r > r_I \qquad (19.121)$$

and

$$\lambda(r) = 10 \qquad\qquad\qquad if\ r \leq r_I \qquad (19.122)$$
$$= 0 \qquad\qquad\qquad if\ r > r_I \qquad (19.123)$$

Dirichlet boundary conditions, corresponding to the analytical solution, are imposed in whole boundary of Ω.
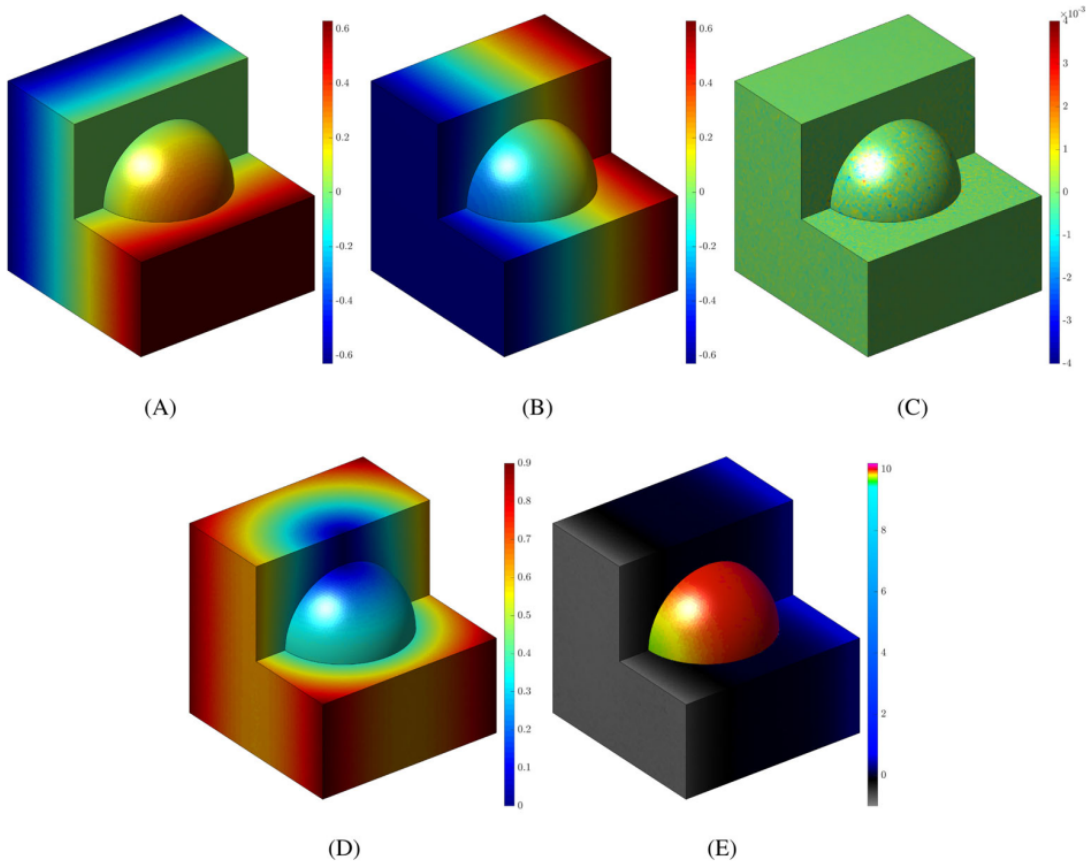
Taken from [1228]. From left to right: u, v, w, $|\vec{v}|$, p, for $\eta_1 = 1$ and $\eta = 10^2$.

19.27 Flow in a square using Stream Functions

I wish to arrive at an analytical formulation for a 2D incompressible flow in the square domain $[-1 : 1] \times [-1 : 1]$ The fluid has constant viscosity $\eta = 1$ and is subject to free slip boundary conditions on all sides. For reasons that will become clear in what follows I postulate the following stream function:

$$\Psi(x, y) = \sin(m\pi x)\sin(n\pi y) \qquad (19.124)$$

We have the velocity being defined as:

$$\vec{v} = (u, v) = \left(\frac{\partial \Psi}{\partial y}, -\frac{\partial \Psi}{\partial x} \right) = (n\pi \sin(m\pi x)\cos(n\pi y), -m\pi \cos(m\pi x)\sin(n\pi y)) \qquad (19.125)$$
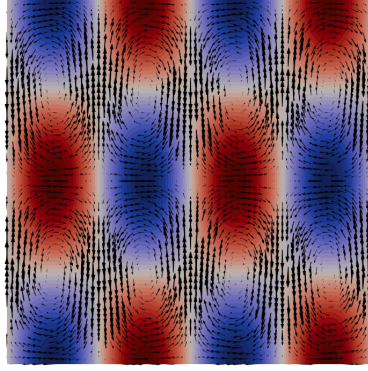
Velocity field for $(m, n) = (2, 1)$

The strain rate components are then:

$$\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} = mn\pi^2 \cos(m\pi x)\cos(n\pi y) \qquad (19.126)$$

$$\dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} = -mn\pi^2 \cos(m\pi x)\cos(n\pi y) \qquad (19.127)$$

$$2\dot{\varepsilon}_{xy} = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \qquad (19.128)$$

$$= \frac{\partial^2 \Psi}{\partial y^2} - \frac{\partial^2 \Psi}{\partial x^2} \qquad (19.129)$$

$$= -n^2 \pi^2 \Psi + m^2 \pi^2 \Psi \qquad (19.130)$$

$$= (m^2 - n^2)\pi^2 \sin(m\pi x)\sin(n\pi y) \qquad (19.131)$$

Note that if $m = n$ the last term is identically zero, which is not desirable (flow is too 'simple') so in what follows we will prefer $m \neq n$.

It is also easy to verify that $u = 0$ on the sides and $v = 0$ at the top and bottom and that the term $\dot{\varepsilon}_{xy}$ is nul on all four sides, thereby garanteeing free slip.

Our choice of stream function yields:

$$\nabla^4 \Psi = \frac{\partial^4 \Psi}{\partial x^4} + \frac{\partial^4 \Psi}{\partial y^4} + 2\frac{\partial^2 \Psi}{\partial x^2 y^2} = \pi^4(m^4 \Psi + n^4 \Psi + 2m^2 n^2 \Psi) = (m^4 + n^4 + 2m^2 n^2)\pi^4 \Psi$$

Let us recall Eq. (??):

$$\vec{\nabla}^4 \Psi = -\frac{\partial \rho g_y}{\partial x} + \frac{\partial \rho g_x}{\partial y} \qquad (19.132)$$

We assume $g_x = 0$ and $g_y = -1$ so that we simply have

$$\frac{\partial \rho}{\partial x} = (m^4 + n^4 + 2m^2 n^2)\pi^4 \Psi = (m^4 + n^4 + 2m^2 n^2)\pi^4 \sin(m\pi x)\sin(n\pi y) \qquad (19.133)$$

so that (assuming the integration constant to be zero):

$$\rho(x, y) = -\frac{m^4 + n^4 + 2m^2 n^2}{m}\pi^3 \cos(m\pi x)\sin(n\pi y)$$

The x-component of the momentum equation is (since $g_x = 0$):

$$-\frac{\partial p}{\partial x} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -\frac{\partial p}{\partial x} - m^2 n\pi^3 \sin(m\pi x)\cos(n\pi y) - n^3 \pi^3 \sin(m\pi x)\cos(n\pi y) = 0$$

so
$$\frac{\partial p}{\partial x} = -(m^2 n + n^3)\pi^3 \sin(m\pi x)\cos(n\pi y)$$

and the pressure field is then (once again neglecting the integration constant):

$$p(x,y) = \frac{m^2 n + n^3}{m}\pi^2 \cos(\pi x)\cos(\pi y)$$

Note that we then have the interesting property that the pressure average over the domain is zero, i.e. $\int p\, dV = 0$.

19.28 One-dimensional advection-diffusion equation

Let us start with the 1D steady advection-diffusion equation:

$$\rho C_p u \frac{dT}{dx} - k\frac{d^2 T}{dx^2} = f \qquad \text{in} \quad [0, L_x] \tag{19.134}$$

with the boundary conditions $T(x=0) = 0$ and $T(x = L_x) = 0$.

The solution to this problem is:

$$T(x) = \frac{f}{u}\left(x - \frac{1 - \exp\frac{ux}{\kappa}}{1 - \exp\frac{u}{\kappa}}\right)$$

or

$$T(x) = \frac{f}{u}\left(x - \frac{1 - \exp\frac{2Pex}{h}}{1 - \exp\frac{2Pe}{h}}\right)$$

where

$$Pe = \frac{uh}{2\kappa} = \frac{uh\rho C_p}{2k}$$

19.29 Annulus with kinematical b.c. - shear flow

Let us consider an annulus domain of inner radius R_1 and outer radius R_2. Boundary conditions are $\vec{v}(r = R_1) = v_1 \vec{e}_\theta$ and $\vec{v}(r = R_2) = v_2 \vec{e}_\theta$. Density is assumed to be zero in the domain.

As seen in Section 19.22, the Laplacian of a vector $\vec{A}$ is given by[6]

$$\nabla^2 \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla \times (\nabla \times \vec{A}) = \begin{pmatrix} \frac{\partial^2 A_r}{\partial r^2} + \frac{1}{r}\frac{\partial A_r}{\partial r} - \frac{1}{r^2}A_r + \frac{1}{r^2}\frac{\partial^2 A_r}{\partial \theta^2} - \frac{2}{r^2}\frac{\partial A_\theta}{\partial \theta} \\ \\ \frac{\partial^2 A_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial A_\theta}{\partial r} - \frac{1}{r^2}A_\theta + \frac{1}{r^2}\frac{\partial^2 A_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial A_r}{\partial \theta} \end{pmatrix}$$

Given the symmetry of the problem and the boundary conditions we know that the solution is as follows:

$$\vec{v}(r,\theta) = v_\theta(r)\vec{e}_\theta$$

Using this velocity field, we can now obtain the pressure field by solving the Stokes equation

$$-\vec{\nabla}p + \eta\vec{\nabla}^2\vec{v} = \vec{0}$$

---

[6]https://en.wikipedia.org/wiki/Vector_Laplacian

since density is zero. Because of symmetry we also expect the pressure to be a function of r only, i.e. $p(r)$. The Stokes equation in polar coordinates then writes:

$$-\frac{\partial p}{\partial r} = 0 \tag{19.135}$$

$$-\frac{1}{r}\underbrace{\frac{\partial p}{\partial \theta}}_{=0} + \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial v_\theta}{\partial r} - \frac{1}{r^2}v_\theta = 0 \tag{19.136}$$

so that the pressure is a constant which we arbitrarily set to zero. We assume $v_\theta(r) = r^\alpha$, and the second equation above becomes:

$$\alpha(\alpha - 1)r^{\alpha-2} + \alpha r^{\alpha-2} - r^{\alpha-2} = 0$$

reducing to $\alpha^2 - 1 = 0$, i.e. $\alpha = \pm 1$, since the above equation must be valid for any value of r. The generic solution then can be written as

$$v_\theta(r) = Ar + \frac{B}{r}$$

Using the b.c. :

$$v_\theta(R_1) = AR_1 + \frac{B}{R_1} = v_1$$

$$v_\theta(R_2) = AR_2 + \frac{B}{R_2} = v_2$$

or,

$$A + \frac{B}{R_1^2} = \frac{v_1}{R_1}$$

$$A + \frac{B}{R_2^2} = \frac{v_2}{R_2}$$

so

$$B = \frac{\frac{v_1}{R_1} - \frac{v_2}{R_2}}{\frac{1}{R_1^2} - \frac{1}{R_2^2}} = R_1 R_2 \frac{v_1 R_2 - v_2 R_1}{R_2^2 - R_1^2}$$

and

$$A = \frac{v_2 R_2 - v_1 R_1}{R_2^2 - R_1^2}$$

$$\boxed{v_\theta(r) = \frac{v_2 R_2 - v_1 R_1}{R_2^2 - R_1^2}r + \frac{R_1 R_2}{r}\frac{v_1 R_2 - v_2 R_1}{R_2^2 - R_1^2}}$$

We can verify that the flow is indeed incompressible:

$$\vec{\nabla} \cdot \vec{v} = \frac{1}{r}\frac{\partial(rv_r)}{\partial r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} = 0$$

since $v_r = 0$ and v_θ does not depend on θ.

Note that we could have used a non-zero density: as long as it does not depend on θ and the gravity points towards the center, it allows for a decoupling of the equations, thereby only contributing to a lithostatic pressure field.

19.30 Generic framework for 3d solution in Cartesian coordinates

We postulate

$$u(x, y, z) = f(x)g'(y)h'(z) \tag{19.137}$$
$$v(x, y, z) = f'(x)g(y)h'(z) \tag{19.138}$$
$$w(x, y, z) = -2f'(x)g'(y)h(z) \tag{19.139}$$

so that the flow is indeed incompressible:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = f'(x)g'(y)h'(z) + f'(x)g'(y)h'(z) - 2f'(x)g'(y)h'(z) = 0$$

The velocity gradient $\boldsymbol{L}(\vec{v})$ is then given by

$$\boldsymbol{L}(\vec{v}) = \begin{pmatrix} f'g'h' & f''gh' & -2f''g'h \\ fg''h' & f'g'h' & -2f'g''h \\ fg'h'' & f'gh'' & -2f'g'h' \end{pmatrix}$$

and the strain rate tensor by:

$$\dot{\varepsilon}(\vec{v}) = \frac{1}{2}(\boldsymbol{L}(\vec{v}) + \boldsymbol{L}(\vec{v})^T) = \frac{1}{2} \begin{pmatrix} 2f'g'h' & (f''g + fg'')h' & fg'h'' - 2f''g'h \\ (f''g + fg'')h' & 2f'g'h' & f'gh'' - 2f'g''h \\ fg'h'' - 2f''g'h & f'gh'' - 2f'g''h & -4f'g'h' \end{pmatrix}$$

We assume for simplicity that $\eta = 1$ so

$$\vec{\nabla} \cdot 2\eta\dot{\varepsilon}(\vec{v}) = \vec{\nabla} \cdot \begin{pmatrix} 2f'g'h' & (f''g + fg'')h' & fg'h'' - 2f''g'h \\ (f''g + fg'')h' & 2f'g'h' & f'gh'' - 2f'g''h \\ fg'h'' - 2f''g'h & f'gh'' - 2f'g''h & -4f'g'h' \end{pmatrix}$$

$$= \begin{pmatrix} 2f''g'h' + (f''g' + fg''')h' + fg'h''' - 2f''g'h' \\ (f'''g + f'g'')h' + 2f'g''h' + f'gh''' - 2f'g''h' \\ f'g'h'' - 2f'''g'h + f'g'h'' - 2f'g'''h - 4f'g'h'' \end{pmatrix}$$

$$= \begin{pmatrix} f''g'h' + fg'''h' + fg'h''' \\ f'''gh' + f'g''h' + f'gh''' \\ -2f'''g'h - 2f'g'''h - 2f'g'h'' \end{pmatrix} \tag{19.140}$$

The Stokes equation then writes

$$\begin{pmatrix} -\partial_x p \\ -\partial_y p \\ -\partial_z p \end{pmatrix} + \begin{pmatrix} f''g'h' + fg'''h' + fg'h''' \\ f'''gh' + f'g''h' + f'gh''' \\ -2f'''g'h - 2f'g'''h - 2f'g'h'' \end{pmatrix} + \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \vec{0}$$

or,

$$\begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} \partial_x p \\ \partial_y p \\ \partial_z p \end{pmatrix} - \begin{pmatrix} f''g'h' + fg'''h' + fg'h''' \\ f'''gh' + f'g''h' + f'gh''' \\ -2f'''g'h - 2f'g'''h - 2f'g'h'' \end{pmatrix}$$

First application   Let us assume

$$f(x) = x(1-x) \qquad g(y) = y(1-y) \qquad h(z) = z(1-z)$$

then

$$f'(x) = 1 - 2x \qquad g'(y) = 1 - 2y \qquad h'(z) = 1 - 2z$$
$$f''(x) = -2 \qquad g''(y) = -2 \qquad h''(z) = -2$$
$$f'''(x) = 0 \qquad g'''(y) = 0 \qquad h'''(z) = 0$$

Which ensures that there is no flow through the boundaries of the unit cube. Then

$$\begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} \partial_x p \\ \partial_y p \\ \partial_z p \end{pmatrix} - \begin{pmatrix} f''g'h' \\ f'g''h' \\ -2f'g'h'' \end{pmatrix}$$

with

$$
\begin{aligned}
u(x,y,z) &= x(1-x)(1-2y)(1-2z) & (19.141) \\
v(x,y,z) &= (1-2x)y(1-y)(1-2z) & (19.142) \\
w(x,y,z) &= -2(1-2x)(1-2y)z(1-z) & (19.143)
\end{aligned}
$$

Finally we postulate

$$p(x,y,z) = -f'g'h' = (2x-1)(2y-1)(2z-1)$$

with $\int_\Omega p(x,y,z)dxdydz = 0$, so

$$\begin{pmatrix} \partial_x p \\ \partial_y p \\ \partial_z p \end{pmatrix} = \begin{pmatrix} -f''g'h' \\ -f'g''h' \\ -f'g'h'' \end{pmatrix} = \begin{pmatrix} 2(2y-1)(2z-1) \\ 2(2x-1)(2z-1) \\ 2(2x-1)(2y-1) \end{pmatrix}$$

Then we find that

$$\vec{b} = \begin{pmatrix} -f''g'h' \\ -f'g''h' \\ -f'g'h'' \end{pmatrix} - \begin{pmatrix} f''g'h' \\ f'g''h' \\ -2f'g'h'' \end{pmatrix} = \begin{pmatrix} -2f''g'h' \\ -2f'g''h' \\ f'g'h'' \end{pmatrix} = \begin{pmatrix} 4(2y-1)(2z-1) \\ 4(2x-1)(2z-1) \\ -2(2x-1)(2y-1) \end{pmatrix}$$

The root mean square velocity is given by

$$\int_\Omega u^2 dV \;\;=\;\; 1/270 \tag{19.144}$$

$$\int_\Omega v^2 dV \;\;=\;\; 1/270 \tag{19.145}$$

$$\int_\Omega w^2 dV \;\;=\;\; 2/135 \tag{19.146}$$

and then

$$\nu_{vrms} = \sqrt{\frac{6}{270}} \simeq 0.1490712$$

This benchmark is implemented in STONE 10 ($Q_1 \times P_0$), STONE 75 (MINI-1 bubble) and STONE 82 (MINI-2 bubbles). The code is available in /mms/mms3D.py.
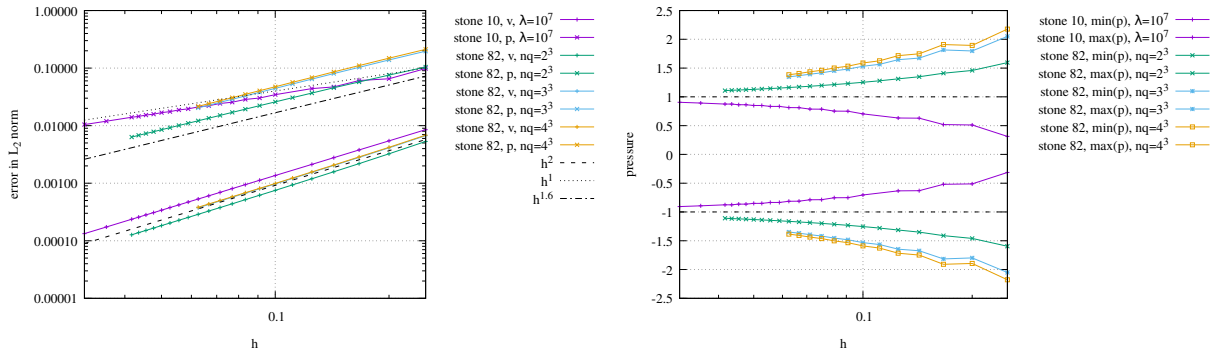
Error convergence and pressure statistics

Second application   Once again, no flow through the boundaries of the unit cube:
$$f(x) = x^2(1-x)^2 \qquad g(y) = y^2(1-y)^2 \qquad h(z) = z^2(1-z)^2$$

then

$$f'(x) = 2x(2x^2 - 3x + 1) \qquad g'(y) = 2y(2y^2 - 3y + 1) \qquad h'(z) = 2z(2z^2 - 3z + 1)$$
$$f''(x) = 2(6x^2 - 6x + 1) \qquad g''(y) = 2(6y^2 - 6y + 1) \qquad h''(z) = 2(6z^2 - 6z + 1)$$
$$f'''(x) = 24x - 12 \qquad g'''(y) = 24y - 12 \qquad h'''(z) = 24z - 12$$

The velocity field is then given by

$$
\begin{aligned}
u(x,y,z) &= f(x)g'(y)h'(z) = 4x^2(1-x)^2y(2y^2 - 3y + 1)z(2z^2 - 3z + 1) \\
v(x,y,z) &= f'(x)g(y)h'(z) = 4x(2x^2 - 3x + 1)y^2(1-y)^2z(2z^2 - 3z + 1) \\
w(x,y,z) &= -2f'(x)g'(y)h(z) = -8x(2x^2 - 3x + 1)y(2y^2 - 3y + 1)z^2(1-z)^2 \quad (19.147)
\end{aligned}
$$

We choose

$$p(x,y,z) = -f'g'h' = -\,2x(2x^2 - 3x + 1)\,2y(2y^2 - 3y + 1)\,2z(2z^2 - 3z + 1)$$

The rhs vector is then

$$
\begin{pmatrix} b_x \\[4pt] b_y \\[4pt] b_z \end{pmatrix}
=
\begin{pmatrix} \partial_x p \\[4pt] \partial_y p \\[4pt] \partial_z p \end{pmatrix}
-
\begin{pmatrix} f''g'h' + fg'''h' + fg'h''' \\[4pt] f'''gh' + f'g''h' + f'gh''' \\[4pt] -2f'''g'h - 2f'g'''h - 2f'g'h'' \end{pmatrix}
$$

$$
= -\begin{pmatrix} f''g'h' \\[4pt] f'g''h' \\[4pt] f'g'h'' \end{pmatrix}
-
\begin{pmatrix} f''g'h' + fg'''h' + fg'h''' \\[4pt] f'''gh' + f'g''h' + f'gh''' \\[4pt] -2f'''g'h - 2f'g'''h - 2f'g'h'' \end{pmatrix}
$$

$$
= -\begin{pmatrix} 2f''g'h' + fg'''h' + fg'h''' \\[4pt] f'''gh' + 2f'g''h' + f'gh''' \\[4pt] -2f'''g'h - 2f'g'''h - f'g'h'' \end{pmatrix}
\qquad (19.148)
$$

Let us look at the root mean square velocity:

$$
\begin{aligned}
\mathbf{v}_{rms}^2 &= \iiint (u^2 + v^2 + w^2)dxdydz \\
&= \iiint u^2 dxdydz + \iiint v^2 dxdydz + \iiint w^2 dxdydz \\
&= \iiint (f(x)g'(y)h'(z))^2 dxdydz + \iiint (f'(x)g(y)h'(z))^2 dxdydz + \iiint (-2f'(x)g'(y)h(z))^2 dxdydz \\
&= \left(\int_0^1 f(x)^2 dx\right)\left(\int_0^1 g'(y)^2 dy\right)\left(\int_0^1 h'(z)^2 dz\right) \\
&+ \left(\int_0^1 f'(x)^2 dx\right)\left(\int_0^1 g(y)^2 dy\right)\left(\int_0^1 h'(z)^2 dz\right) \\
&+ 4\left(\int_0^1 f'(x)^2 dx\right)\left(\int_0^1 g'(y)^2 dy\right)\left(\int_0^1 h(z)^2 dz\right)
\end{aligned}
$$

$$(19.149)$$

Since

$$\int_0^1 f(x)^2 dx = \int_0^1 g(y)^2 dy = \int_0^1 h(z)^2 dz$$

and

$$\int_0^1 g'(y)^2 dy = \int_0^1 f'(x)^2 dx = \int_0^1 h'(z)^2 dz$$

we only have 2 integrals to compute and using WolframAlpha we find:

$$
\begin{aligned}
\left(\int_0^1 f(x)^2 dx \right) &= \frac{1}{630} \\
\left(\int_0^1 f'(x)^2 dx \right) &= \frac{2}{105}
\end{aligned}
$$

so that

$$
\begin{aligned}
v_{rms}^2 &= \frac{1}{630}\frac{2}{105}\frac{2}{105} + \frac{2}{105}\frac{1}{630}\frac{2}{105} + 4\frac{1}{630}\frac{2}{105}\frac{1}{630} \\
&= 6\frac{1}{630}\frac{2}{105}\frac{2}{105} \\
&= \frac{1}{105}\frac{2}{105}\frac{2}{105}
\end{aligned}
\tag{19.150}
$$

In the end:

$$v_{rms} \simeq 0.00185885728$$

This benchmark is carried out in Stone 10.

19.31   2d Analytical benchmark XII

This is presented in Soulaimani, Fortin, Ouellet, Dhatt, and Bertrand [1261] (1987). The velocity field is given by

$$\vec{v}(x,y) = (x^3, -3x^2 y)$$

and the pressure is

$$p(x,y) = x^3 + y^3 - 1/2$$

so that, assuming that the viscosity is 1, the body force is:

$$\vec{b} = (-6x + 3x^2, 6y + 3y^2)$$

Note that I have added the $-1/2$ term to the pressure so that $\int\int p\, dx dy = 0$. The root mean square velocity over a unit square is

$$v_{rms} = \sqrt{\int_0^1 \int_0^1 (u^2 + v^2) dx dy} = \sqrt{\int_0^1 \int_0^1 (x^6 + 9x^4 y^2) dx dy} = \sqrt{\frac{1}{7} + 9\frac{1}{5}\frac{1}{3}} = \sqrt{\frac{26}{35}} \simeq 0.861892$$

The strain rate tensor terms are

$$
\begin{aligned}
\dot{\varepsilon}_{xx} &= 3x^2 \\
\dot{\varepsilon}_{yy} &= -3x^2 \\
\dot{\varepsilon}_{xy} &= -3xy
\end{aligned}
$$

Another one mentioned in the paper:

$$\vec{v}(x,y) = (x^2, -2xy) \qquad p(x,y) = 0 \qquad \vec{b} = (-2, 0)$$

19.32   2d analytical benchmark from Burman & Hansbo (2006)

This is presented in Burman and Hansbo [195] (2006) and Burman and Hansbo [196] (2007) and apparently originates in Norburn and Silvester [1002] (1998).

The velocity and pressure fields are given in the unit square by

$$
\begin{aligned}
u(x,y) &= 20xy^3 \\
v(x,y) &= 5x^4 - 5y^4 \\
p(x,y) &= 60x^2y - 20y^3 - 5
\end{aligned}
\tag{19.151}
$$

with

$$
\begin{aligned}
\frac{\partial u}{\partial x} &= 20y^3 \\
\frac{\partial u}{\partial y} &= 60xy^2 \\
\frac{\partial v}{\partial x} &= 20x^3 \\
\frac{\partial v}{\partial y} &= -20y^3
\end{aligned}
\tag{19.152}
$$
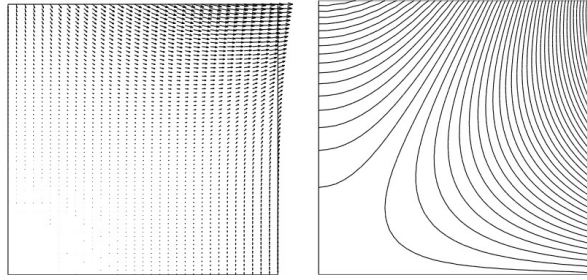
Taken from Burman and Hansbo [195]. Left is velocity, right is pressure.

The flow is incompressible:

$$
div(\vec{v}) = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0
$$

Then the strain rate tensor is given by

$$
\dot{\varepsilon}(\vec{v}) = \begin{pmatrix} 20y^3 & 30xy^2 + 10x^3 \\ 30xy^2 + 10x^3 & -20y^3 \end{pmatrix}
$$

Assuming the viscosity $\eta = 1$, then the full stress tensor is given by

$$
\begin{aligned}
\boldsymbol{\sigma} &= -p\mathbf{1} + 2\eta\dot{\varepsilon}(\vec{v}) \\
&= \begin{pmatrix} -60x^2y + 20y^3 + 5 + 40y^3 & 60xy^2 + 20x^3 \\ 60xy^2 + 20x^3 & -60x^2y + 20y^3 + 5 - 40y^3 \end{pmatrix} \\
&= \begin{pmatrix} -60x^2y + 60y^3 + 5 & 60xy^2 + 20x^3 \\ 60xy^2 + 20x^3 & -60x^2y - 20y^3 + 5 \end{pmatrix}
\end{aligned}
\tag{19.153}
$$

finally

$$
\vec{b} = -\vec{\nabla} \cdot \boldsymbol{\sigma} = \begin{pmatrix} -120xy + 120xy \\ 60y^2 + 60x^2 - 60x^2 - 60y^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
$$

This is particularly convenient...

It is implemented in STONE 14, 18 and 115.

19.33 2d analytical benchmark from Cioncolini & Boffi (2019)

This is Test #1 presented in Cioncolini and Boffi [276] (2019). The domain is a unit square and the solution is given by

$$
\begin{aligned}
u(x,y) &= x^2(1-x)^2 2y(1-y)(2y-1) \\
v(x,y) &= y^2(1-y)^2 2x(1-x)(1-2x) \\
p(x,y) &= x(1-x)(1-y) - 1/12 \\
b_x &= -\eta\left\{4y(1-y)(2y-1)[(1-2x)^2 - 2x(1-x)] + 12x^2(1-x)^2(1-2y)\right\} + (1-2x)(1-y) \quad (19.154) \\
b_y &= \quad (19.155)
\end{aligned}
$$

with no-slip boundary conditions on all sides. Note that the veloity solution is close to the mms of Section ??

19.34 Poisson equation on 3d shell

This benchmark is presented in Phillips *et al.* [1058]. Inner radius is 1, outer radius is 3. The right hand side term of the Poisson equation is given by:

$$
\begin{aligned}
f(r,\theta,\phi) &= \frac{\sin^2\theta}{R^2}\left[(\cos\phi - \sin\phi)(20\sin^2\theta - 15) - \sin 2\phi(10\sin^2\theta - 6)\right] \\
&\times \left[\left(\frac{r}{R_{inner}}\right)^2 - 1\right]\left[\left(\frac{r}{R_{outer}}\right)^2 - 1\right] \\
&+ \sin^4\theta\left[\cos\phi - \sin\phi - \frac{1}{2}\sin 2\phi\right]\left[\frac{20r^2}{R_{inner}^2 R_{outer}^2} - 6\left(\frac{1}{R_{inner}^2} + \frac{1}{R_{outer}^2}\right)\right] \quad (19.156)
\end{aligned}
$$

Note: what is R here??

The solution is then

$$
T(r,\theta,\phi) = \sin^4\theta\left[\cos\phi - \sin\phi - \frac{1}{2}\sin 2\phi\right]\left[\left(\frac{r}{R_{inner}}\right)^2 - 1\right]\left[\left(\frac{r}{R_{outer}}\right)^2 - 1\right]
$$

This expression is used to generate the Dirichlet boundary conditions on the inner and outer surfaces.

19.35 Manufactured solution 1 in John, Linke, Merdon, Neilan, and Rebholz [693]

It is presented in Example 1.1 of [693] and meant to be a peculiar case where the velocity solution is exactly zero. The viscosity is 1, the domain is a unit square, no-slip boundary conditions are prescribed everywhere. The buoyancy force is given by $\vec{b} = (0, Ra(1 - y + 3y^2))$ where $Ra > 0$ is a parameter. The flow is incompressible and the analytical pressure solution is given by $p = Ra(y^3 - y^2/2 + y - 7/12)$.

19.36 Manufactured solution 2 in John, Linke, Merdon, Neilan, and Rebholz [693]

This benchmark comes from John *et al.* [693]. The domain is once again the unit square. The velocity field has the form of a large vortex. Note that velocity field is actually the same velocity field as in the Donea & Huerta benchmark above (albeit multiplied by a factor 100).

$$
\begin{align}
u(x,y) &= 200x^2(1-x)^2 y(1-y)(1-2y) \tag{19.157}\\
v(x,y) &= -200x(1-x)(1-2x)y^2(1-y)^2 \tag{19.158}\\
p(x,y) &= 10\left[(x-1/2)^3 y^2 + (1-x)^3(y-1/2)^3\right] \tag{19.159}
\end{align}
$$
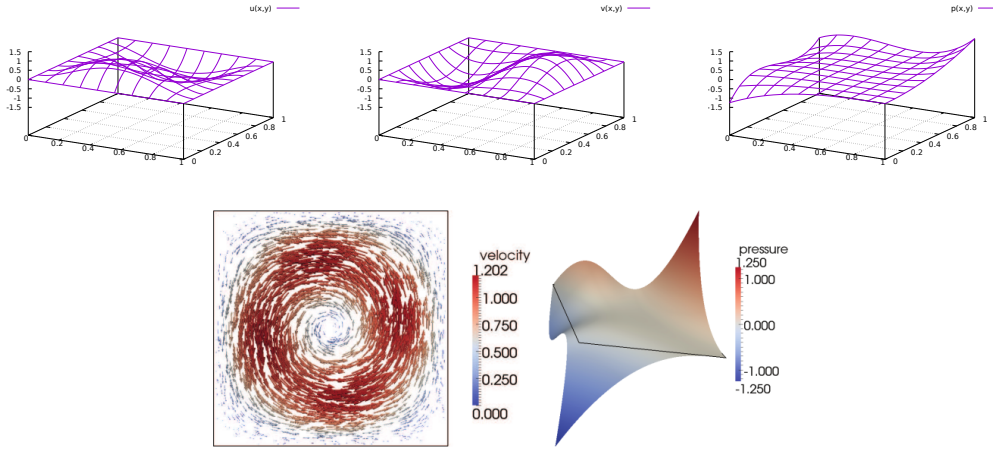
Taken from John, Linke, Merdon, Neilan, and Rebholz [693] (2017).

$$
\begin{align}
\dot{\varepsilon}_{xx} = \frac{\partial u}{\partial x} &= -400(1-x)x(2x-1)(y-1)y(2y-1) \tag{19.160}\\
\frac{\partial u}{\partial y} &= 200(1-x)^2 x^2(6y^2-6y+1) \tag{19.161}\\
\frac{\partial v}{\partial x} &= -200(6x^2-6x+1)(1-y)^2 y^2 \tag{19.162}\\
\dot{\varepsilon}_{yy} = \frac{\partial v}{\partial y} &= 400(x-1)x(2x-1)(1-y)y(2y-1) \tag{19.163}
\end{align}
$$

so that

$$
\begin{align}
\dot{\varepsilon}_{xy} &= \frac{1}{2}\left[200(1-x)^2 x^2(6y^2-6y+1) - 200(6x^2-6x+1)(1-y)^2 y^2\right]\\
&= 100(1-x)^2 x^2(6y^2-6y+1) - 100(6x^2-6x+1)(1-y)^2 y^2 \tag{19.164}
\end{align}
$$

Also

$$\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} = 400(6x^2 - 6x + 1)y(2y^2 - 3y + 1)$$

$$\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} = 200(-2x^2(1-x)(6y^2 - 6y + 1) + 2x(1-x)^2(6y^2 - 6y + 1) - 6(2x-1)(1-y)^2y^2)$$

$$= 100(-2x^2(1-x)(6y^2 - 6y + 1) + 2x(1-x)^2(6y^2 - 6y + 1) - 6(2x-1)(1-y)^2y^2)$$

$$\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} = 400(6x^2 - 6x + 1)(1-y)y^2 + 200(1-x)^2x^2(12y - 6) - 400(6x^2 - 6x + 1)(1-y)^2y$$

$$\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} = -400x(2x^2 - 3x + 1)(6y^2 - 6y + 1) \tag{19.165}$$

$$\frac{\partial p}{\partial x} = 30(x - 1/2)^2y^2 - 30(1-x)^2(y - 1/2)^3 \tag{19.166}$$

$$\frac{\partial p}{\partial y} = 20(x - 1/2)^3y + 30(1-x)^3(y - 1/2)^2 \tag{19.167}$$

From $\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{b} = \vec{0}$ we can obtain the rhs as follows:

$$\vec{b} = -\vec{\nabla} \cdot \boldsymbol{\sigma}$$

$$= \vec{\nabla}p - \vec{\nabla} \cdot \boldsymbol{s}$$

$$= \vec{\nabla}p - \vec{\nabla} \cdot (2\eta\dot{\boldsymbol{\varepsilon}}) \tag{19.168}$$

Assuming $\eta = 1$ we arrive at:

$$b_x = \frac{\partial p}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xx}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial y} \tag{19.169}$$

$$b_y = \frac{\partial p}{\partial y} - 2\frac{\partial \dot{\varepsilon}_{xy}}{\partial x} - 2\frac{\partial \dot{\varepsilon}_{yy}}{\partial y} \tag{19.170}$$

All the necessary functions to do this benchmark are in `mms/jolm17.py`:

```
# functions for the John et al (2017) manufactured solution

def u_th(x,y):
    return 200*x**2*(1-x)**2*y*(1-y)*(1-2*y)

def v_th(x,y):
    return -200*x*(1-x)*(1-2*x)*y**2*(1-y)**2

def p_th(x,y):
    return 10*( (x-1./2.)**3*y**2+(1-x)**3*(y-1./2.)**3 )

def dpdx_th(x,y):
    return 30*(x-1./2.)**2*y**2-30*(1-x)**2*(y-1./2.)**3

def dpdy_th(x,y):
    return 20*(x-1./2.)**3*y + 30*(1-x)**3*(y-1./2.)**2

def exx_th(x,y):
    return -400*(1-x)*x*(2*x-1)*(y-1)*y*(2*y-1)

def exy_th(x,y):
    return 100*(1-x)**2*x**2*(6*y**2-6*y+1)-100*(6*x**2-6*x+1)*(1-y)**2*y**2
```

```python
def eyy_th(x,y):
    return  400*(x-1)*x*(2*x-1)*(1-y)*y*(2*y-1)

def dexxdx(x,y):
    return  400*(6*x**2-6*x+1)*y*(2*y**2-3*y+1)

def dexydx(x,y):
    return  100*(-2*x**2*(1-x)*(6*y**2-6*y+1) + 2*x*(1-x)**2*(6*y**2-6*y+1)  -6*(2*x-1)*(1-y)**2*y**2)

def dexydy(x,y):
    return  200*(6*x**2-6*x+1)*(1-y)*y**2 + 100*(1-x)**2*x**2*(12*y-6)  -200*(6*x**2-6*x+1)*(1-y)**2*y

def deyydy(x,y):
    return  -400*x*(2*x**2-3*x+1)*(6*y**2-6*y+1)

def bx(x,y):
    return  dpdx_th(x,y)-2*dexxdx(x,y)-2*dexydy(x,y)

def by(x,y):
    return  dpdy_th(x,y)-2*dexydx(x,y)-2*deyydy(x,y)
```
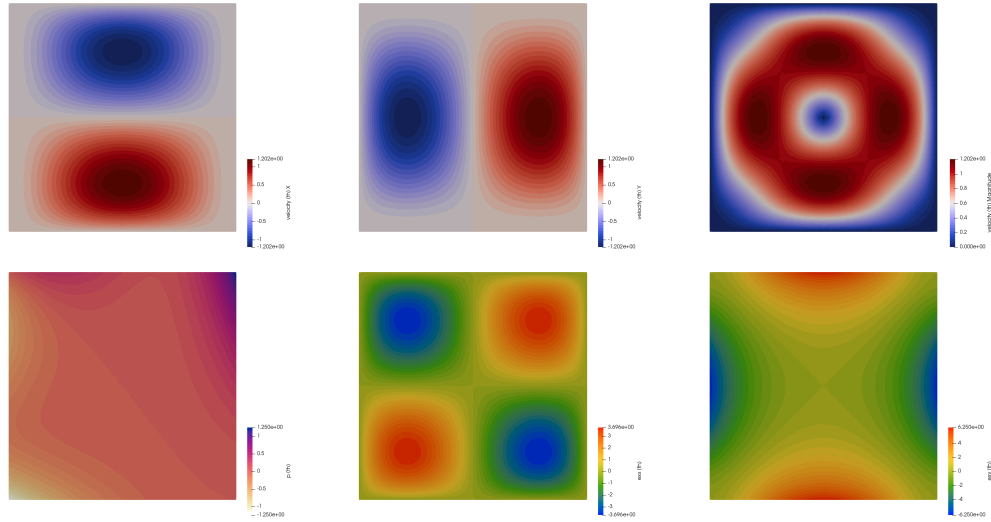

At some point I had also written this about he very same benchmark. The velocity is given by

$$
\begin{aligned}
\nu_x &= 200x^2(1-x)^2 y(1-y)(1-2y) = 100a(x)a'(y) \\
\nu_y &= -200x(1-x)(1-2x)y^2(1-y)^2 = -100a'(x)a(y)
\end{aligned}
$$

with

$$
\begin{aligned}
a(x) &= x^2(1-x)^2 \\
a'(x) &= 2x(1-x)^2 - 2x^2(1-x) = 2x(1-x)(1-2x) \\
a''(x) &= 2(1-6x+6x^2) \\
a'''(x) &= 24x - 12
\end{aligned}
$$

We can now compute the components of the strain rate tensor:

$$
\begin{aligned}
\dot{\varepsilon}_{xx} = \frac{\partial \mathbf{v}_x}{\partial x} &= \frac{\partial(100a(x)a'(y))}{\partial x} = 100a'(x)a'(y) \\
\dot{\varepsilon}_{yy} = \frac{\partial \mathbf{v}_y}{\partial y} &= \frac{\partial(-100a'(x)a(y))}{\partial y} = -100a'(x)a'(y) \\
\dot{\varepsilon}_{xy} = \dot{\varepsilon}_{yx} = \frac{1}{2}\left(\frac{\partial \mathbf{v}_x}{\partial y} + \frac{\partial \mathbf{v}_y}{\partial x}\right) &= \frac{1}{2}\left(100a(x)a''(y) - 100a''(x)a(y)\right) = 50\left(a(x)a''(y) - a''(x)a(y)\right)
\end{aligned}
$$

The momentum conservation equation is given by

$$
\begin{aligned}
-\partial_x p + \partial_x(2\eta \dot{\varepsilon}_{xx}) + \partial_y(2\eta \dot{\varepsilon}_{xy}) + b_x &= 0 \\
-\partial_y p + \partial_x(2\eta \dot{\varepsilon}_{xy}) + \partial_y(2\eta \dot{\varepsilon}_{yy}) + b_y &= 0
\end{aligned}
$$

Then

$$
\begin{aligned}
b_x &= \partial_x p - \partial_x(2\eta \dot{\varepsilon}_{xx}) - \partial_y(2\eta \dot{\varepsilon}_{xy}) \\
&= \partial_x p - \partial_x[2\eta 100 a'(x)a'(y)] - \partial_y[2\eta 50\left(a(x)a''(y) - a''(x)a(y)\right)] \\
&= \partial_x p - 200\eta a''(x)a'(y) - 100\eta[a(x)a'''(y) - a''(x)a'(y)] \\
&= \partial_x p - 100\eta a''(x)a'(y) - 100\eta a(x)a'''(y) \\
&= \partial_x p - 100\eta[a''(x)a'(y) + a(x)a'''(y)]
\end{aligned}
$$

$$
\begin{aligned}
b_y &= \partial_y p - \partial_x(2\eta \dot{\varepsilon}_{xy}) - \partial_y(2\eta \dot{\varepsilon}_{yy}) \\
&= \partial_y p - \partial_x[2\eta 50\left(a(x)a''(y) - a''(x)a(y)\right)] + \partial_y 2\eta 100 a'(x)a'(y) \\
&= \partial_y p - 100\eta(a'(x)a''(y) - a'''(x)a(y)) + 200\eta a'(x)a''(y) \\
&= \partial_y p + 100\eta a'(x)a''(y) + 100\eta a'''(x)a(y) \\
&= \partial_y p + 100\eta[a'(x)a''(y) + a'''(x)a(y)]
\end{aligned}
$$

with

$$
\begin{aligned}
p(x,y) &= 10\left[\left(x - \frac{1}{2}\right)^3 y^2 + (1-x)^3\left(y - \frac{1}{2}\right)^3\right] \\
\frac{\partial p}{\partial x} &= 10\left[3\left(x - \frac{1}{2}\right)^2 y^2 - 3(1-x)^2\left(y - \frac{1}{2}\right)^3\right] \\
\frac{\partial p}{\partial y} &= 10\left[\left(x - \frac{1}{2}\right)^3 2y + (1-x)^3 3\left(y - \frac{1}{2}\right)^2\right]
\end{aligned}
$$

See STONE 104.

19.37 Channel flow with Herschel-Bulkley rheology

We start from the following formulation for the Herschel-Bulkley rheology:

$$
\eta_{HB} = \begin{cases} \eta_0 & \dot{\varepsilon}_e \leq \dot{\varepsilon}_0 \\ K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_e} & \dot{\varepsilon}_e \geq \dot{\varepsilon}_0 \end{cases}
$$

and the limiting viscosity η_0 is such that

$$\eta_0 = K\dot{\varepsilon}_0^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_0}$$

We consider a two-dimensional channel in the x, y plane. The walls are at $y = 0$ and $y = H$ with no-slip boundary conditions. In the absence of gravity, the Stokes equation simplify to

$$-\frac{\partial p}{\partial x} + \frac{\partial}{\partial y}(2\eta_{HB}\dot{\varepsilon}_{xy}) = 0 \qquad \text{and} \qquad \dot{\varepsilon}_{xy} = \frac{1}{2}\frac{\partial u}{\partial y} \tag{19.171}$$

where we assume the velocity $\vec{v} = (u(y), 0)$. It then follows that

$$\dot{\varepsilon}_e = \sqrt{\mathcal{I}_2(\dot{\boldsymbol{\varepsilon}})} = \sqrt{\frac{1}{2}\dot{\boldsymbol{\varepsilon}} : \dot{\boldsymbol{\varepsilon}}} = \sqrt{\frac{1}{2}[(\dot{\varepsilon}_{xx})^2 + (\dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{zz})^2] + (\dot{\varepsilon}_{xy})^2 + (\dot{\varepsilon}_{xz})^2 + (\dot{\varepsilon}_{yz})^2} = \sqrt{\dot{\varepsilon}_{xy}^2} = \left|\frac{1}{2}\frac{\partial u}{\partial y}\right|$$

In the case of a Newtonian fluid, the analytical solution is known and the velocity profile is a parabola with zero velocity on the walls and maximum velocity in the middle. Although the rheology of the fluid is non-linear we assume that a similar velocity profile is expected (although not described by a parabola). We then expect three zones (and we assume that the fluid flows from left to right):

- In the middle, where it is expected that $\frac{\partial u}{\partial y} = 0$ (at least in one point) because of symmetry. We also therefore expect $\dot{\varepsilon}_e \leq \dot{\varepsilon}_0$ in this region so that $\eta_{HB} = \eta_0$. How thick this region is will be determined later.

  Eq. (19.171) must then be solved

$$\frac{\partial p}{\partial x} = \frac{\partial}{\partial y}\left(2\eta_{HB}\frac{1}{2}\frac{\partial u}{\partial y}\right) = \eta_0 \frac{\partial^2 u}{\partial y^2} \tag{19.172}$$

  Let us call $\Pi = \frac{\partial p}{\partial x} < 0$, then we must solve:

$$\frac{\partial^2 u}{\partial y^2} = \frac{\Pi}{\eta_0}$$

  The solution is then of the form

$$\boxed{u(y)|_{mid} = \frac{1}{2}\frac{\Pi}{\eta_0}y^2 + 2ay + b}$$

  and

$$\boxed{\dot{\varepsilon}_{xy}|_{mid} = \frac{1}{2}\frac{\Pi}{\eta_0}y + a}$$

  We will determine a and b later.

- Near the bottom wall, with $\frac{\partial u}{\partial y} > 0$ so that $\dot{\varepsilon}_e = +\frac{1}{2}\left(\frac{\partial u}{\partial y}\right)$ and $\dot{\varepsilon}_e \geq \dot{\varepsilon}_0$. We solve Eq. (19.171) again, this time with the non-linear formulation of the viscosity:

$$
\begin{aligned}
\frac{\partial p}{\partial x} &= \frac{\partial}{\partial y}\left(2\eta_{HB}\frac{1}{2}\frac{\partial u}{\partial y}\right) \\
&= 2\frac{\partial}{\partial y}\left[\left(K\dot{\varepsilon}_e^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_e}\right)\frac{1}{2}\frac{\partial u}{\partial y}\right] \\
&= 2\frac{\partial}{\partial y}\left[\left(K\left|\frac{1}{2}\frac{\partial u}{\partial y}\right|^{n-1} + \tau_0\left|\frac{1}{2}\frac{\partial u}{\partial y}\right|^{-1}\right)\frac{1}{2}\frac{\partial u}{\partial y}\right] \\
&= 2\frac{\partial}{\partial y}\left[K\left(\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0\right] \tag{19.173}
\end{aligned}
$$

1009

We then must solve:

$$\frac{\partial}{\partial y}\left[K\left(\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0\right] = \frac{\Pi}{2}$$

$$K\left(\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0 = \frac{\Pi}{2}y + c$$

$$\left(\frac{1}{2}\frac{\partial u}{\partial y}\right)^n = \frac{1}{K}(\frac{\Pi}{2}y + c - \tau_0)$$

or,

$$\boxed{\dot\varepsilon_{xy}|_{bot} = \frac{1}{2}\frac{\partial u}{\partial y} = \left(\frac{1}{K}(\Pi_2 y + c - \tau_0)\right)^{1/n}}$$

so

$$\boxed{u(y)|_{bot} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(\Pi_2 y + c - \tau_0)\right)^{1+1/n} + d}$$

where $\Pi_2 = \Pi/2$

- Near the top wall, with $\frac{\partial u}{\partial y} < 0$ so that $\dot\varepsilon_e = -\frac{1}{2}\left(\frac{\partial u}{\partial y}\right)$ and $\dot\varepsilon_e \geq \dot\varepsilon_0$. We solve yet again Eq. (19.171):

$$\begin{aligned}
\frac{\partial p}{\partial x} &= \frac{\partial}{\partial y}\left(2\eta_{HB}\frac{1}{2}\frac{\partial u}{\partial y}\right) \\
&= 2\frac{\partial}{\partial y}\left[\left(K\dot\varepsilon_e^{n-1} + \frac{\tau_0}{\dot\varepsilon_e}\right)\frac{1}{2}\frac{\partial u}{\partial y}\right] \\
&= 2\frac{\partial}{\partial y}\left[\left(K\left|\frac{1}{2}\frac{\partial u}{\partial y}\right|^{n-1} + \tau_0\left|\frac{1}{2}\frac{\partial u}{\partial y}\right|^{-1}\right)\frac{1}{2}\frac{\partial u}{\partial y}\right] \\
&= -2\frac{\partial}{\partial y}\left[\left(K\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^{n-1} + \tau_0\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^{-1}\right)\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)\right] \\
&= -2\frac{\partial}{\partial y}\left[K\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0\right] \qquad (19.174)
\end{aligned}$$

We then must solve:

$$-\frac{\partial}{\partial y}\left[K\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0\right] = \frac{\Pi}{2}$$

$$K\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^n + \tau_0 = -\frac{\Pi}{2}y + e$$

$$\left(-\frac{1}{2}\frac{\partial u}{\partial y}\right)^n = \frac{1}{K}(-\frac{\Pi}{2}y + e - \tau_0)$$

which yields

$$\boxed{\dot\varepsilon_{xy}|_{top} = -\left(\frac{1}{K}(-\Pi_2 y + e - \tau_0)\right)^{1/n}}$$

$$\boxed{u(y)|_{top} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(-\Pi_2 y + e - \tau_0)\right)^{1+1/n} + f}$$

We have 6 integration constants a, b, c, d, e, f and 6 additional constraints from continuity or boundary conditions:

$$(1) \quad u(0) = 0 \text{ boundary condition} \qquad (19.175)$$
$$(2) \quad u(H) = 0 \text{ boundary condition} \qquad (19.176)$$
$$(3) \quad u(y_1) \text{ must be continuous} \qquad (19.177)$$
$$(4) \quad u(y_2) \text{ must be continuous} \qquad (19.178)$$
$$(5) \quad \dot{\varepsilon}_{xy}(y_1) \text{ must be continuous} \qquad (19.179)$$
$$(6) \quad \dot{\varepsilon}_{xy}(y_2) \text{ must be continuous} \qquad (19.180)$$

Using symmetry to compute a Because of symmetry, we expect $y_1 = H/2 - \delta$ and $y_2 = H/2 + \delta$ with $\delta \neq 0$ (i.e. $y_1 \neq y_2$) and we expect $u(y_1) = u(y_2)$ so that

$$u(y_1)|_{mid} = \frac{1}{2}\frac{\Pi}{\eta_0}y_1^2 + 2ay_1 + b = \frac{1}{2}\frac{\Pi}{\eta_0}y_2^2 + 2ay_2 + b = u(y_2)|_{mid}$$

or,

$$\frac{1}{2}\frac{\Pi}{\eta_0}(y_1^2 - y_2^2) + 2a(y_1 - y_2) = 0$$

$$\frac{1}{2}\frac{\Pi}{\eta_0}(y_1 - y_2)(y_1 + y_2) + 2a(y_1 - y_2) = 0$$

$$\frac{1}{2}\frac{\Pi}{\eta_0}(y_1 + y_2) + 2a = 0$$

$$\frac{1}{2}\frac{\Pi}{\eta_0}H + 2a = 0$$

and finally we obtain a:

$$\boxed{a = -\frac{1}{4}\frac{\Pi}{\eta_0}H}$$

Note that we could have obtained the same thing by enforcing that the strain rate at y_1 and y_2 are the opposite of one another. It then follows:

$$\boxed{u(y)|_{mid} = \frac{1}{2}\frac{\Pi}{\eta_0}y^2 - 2\frac{1}{4}\frac{\Pi}{\eta_0}Hy + b = \frac{1}{2}\frac{\Pi}{\eta_0}(y^2 - yH) + b = \frac{\Pi_2}{\eta_0}(y^2 - yH) + b}$$

and

$$\boxed{\dot{\varepsilon}_{xy}|_{mid} = \frac{1}{2}\frac{\Pi}{\eta_0}y - \frac{1}{4}\frac{\Pi}{\eta_0}H = \frac{1}{2}\frac{\Pi}{\eta_0}\left(y - \frac{H}{2}\right) = \frac{\Pi_2}{\eta_0}\left(y - \frac{H}{2}\right)}$$

Because of the parabola-like flow profile, we expect the strain rate to be zero in the middle $y = H/2$, and positive for $z_1 < y < H/2$ and negative for $H/2 < y < z_2$, which is indeed what we recover ($\Pi < 0$).

Using bottom boundary condition to obtain d

$$u(y = 0)|_{bot} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(c - \tau_0)\right)^{1+1/n} + d = 0$$

so

$$d = -2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(c - \tau_0)\right)^{1+1/n}$$

and then

$$\boxed{u(y)|_{bot} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{1}{K}(\Pi y + c - \tau_0)\right)^{1+1/n} - \left(\frac{1}{K}(c - \tau_0)\right)^{1+1/n}\right]}$$

1011

Using top boundary condition to obtain f

$$u(y = H)|_{top} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(-\Pi_2 H + e - \tau_0)\right)^{1+1/n} + f = 0$$

so

$$f = -2\frac{n}{n+1}\frac{K}{\Pi_2}\left(\frac{1}{K}(-\Pi_2 H + e - \tau_0)\right)^{1+1/n}$$

and then

$$\boxed{u(y)|_{top} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{1}{K}(-\Pi_2 y + e - \tau_0)\right)^{1+1/n} - \left(\frac{1}{K}(-\Pi_2 H + e - \tau_0)\right)^{1+1/n}\right]}$$

computing δ   The coordinates of the transitions y_1 and y_2 are the location where the strain rate $\dot{\varepsilon}_e$ reaches $\dot{\varepsilon}_0$. In other words:

$$\dot{\varepsilon}_e|_{mid}(y = y_1) = \dot{\varepsilon}_{xy}|_{mid}(y = y_1) = \frac{1}{2}\frac{\Pi}{\eta_0}\left(y_1 - \frac{H}{2}\right) = \frac{1}{2}\frac{\Pi}{\eta_0}\left(\frac{H}{2} - \delta - \frac{H}{2}\right) = -\frac{1}{2}\frac{\Pi}{\eta_0}\delta = \dot{\varepsilon}_0$$

or,

$$\delta = -\frac{2\dot{\varepsilon}_0 \eta_0}{\Pi}$$

Since $\Pi < 0$ it adds up and $\delta > 0$. We can also write

$$\boxed{\delta = \frac{2\dot{\varepsilon}_0 \eta_0}{|\Pi|}}$$

and we will use throughout what follows:

$$\dot{\varepsilon}_0 = -\frac{1}{2}\frac{\Pi}{\eta_0}\delta$$

Using strain rate continuity at y_1 to compute c

$$\left(\frac{1}{K}(\Pi_2 y_1 + c - \tau_0)\right)^{1/n} = \frac{1}{2}\frac{\Pi}{\eta_0}\left(y_1 - \frac{H}{2}\right) = -\frac{1}{2}\frac{\Pi}{\eta_0}\delta$$

$$\Pi_2 y_1 + c - \tau_0 = K\left(-\frac{1}{2}\frac{\Pi}{\eta_0}\delta\right)^n$$

$$c = K\left(-\frac{1}{2}\frac{\Pi}{\eta_0}\delta\right)^n + \tau_0 - \Pi_2 y_1$$

$$\boxed{c = K\dot{\varepsilon}_0^n + \tau_0 - \Pi_2 y_1}$$

$$u(y)|_{bot} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{1}{K}(\Pi_2 y + c - \tau_0)\right)^{1/n+1} - \left(\frac{1}{K}(c - \tau_0)\right)^{1/n+1}\right]$$

$$= 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{1}{K}(\Pi_2(y - y_1) + K\dot{\varepsilon}_0^n)\right)^{1/n+1} - \left(\frac{1}{K}(K\dot{\varepsilon}_0^n - \Pi_2 y_1)\right)^{1/n+1}\right]$$

$$= 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{\Pi_2}{K}(y - y_1) + \dot{\varepsilon}_0^n\right)^{1/n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right]$$

$$\dot{\varepsilon}_{xy}|_{bot} = \left(\frac{1}{K}(\Pi_2 y + c - \tau_0)\right)^{1/n}$$

$$= \left(\frac{\Pi_2}{K}(y - y_1) + \dot{\varepsilon}_0^n\right)^{1/n}$$

Using strain rate continuity at y_2 to compute e

$$-\left(\frac{1}{K}(-\Pi_2 y_2 + e - \tau_0)\right)^{1/n} = \frac{1}{2}\frac{\Pi}{\eta_0}\left(y_2 - \frac{H}{2}\right) = \frac{1}{2}\frac{\Pi}{\eta_0}\delta$$

$$-\Pi_2 y_2 + e - \tau_0 = K\left(-\frac{1}{2}\frac{\Pi}{\eta_0}\delta\right)^n$$

$$e = K\left(-\frac{1}{2}\frac{\Pi}{\eta_0}\delta\right)^n + \tau_0 + \Pi_2 y_2$$

$$\boxed{e = K\dot{\varepsilon}_0^n + \tau_0 + \Pi_2 y_2}$$

$$u(y)|_{top} = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{1}{K}(-\Pi_2 y + e - \tau_0)\right)^{1/n+1} - \left(\frac{1}{K}(-\Pi_2 H + e - \tau_0)\right)^{1/n+1}\right]$$

$$= 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(-\frac{\Pi_2}{K}(y - y_2) + \dot{\varepsilon}_0^n\right)^{1/n+1} - \left(-\frac{\Pi_2}{K}(H - y_2) + \dot{\varepsilon}_0^n\right)^{1/n+1}\right]$$

$$\dot{\varepsilon}_{xy}|_{top} = -\left(\frac{1}{K}(-\Pi_2 y + e - \tau_0)\right)^{1/n}$$

$$= -\left(-\frac{\Pi_2}{K}(y - y_2) + \dot{\varepsilon}_0^n\right)^{1/n}$$

Using velocity continuity to compute b   We use $u(y_1)|_{bot} = u(y_1)|_{mid}$:

$$2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{\Pi_2}{K}(y_1 - y_1) + \dot{\varepsilon}_0^n\right)^{1/n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right] = \frac{\Pi_2}{\eta_0}(y_1^2 - y_1 H) + b$$

$$2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\dot{\varepsilon}_0^{n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right] = \frac{\Pi_2}{\eta_0}y_1(y_1 - H) + b$$

so

$$b = 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\dot{\varepsilon}_0^{n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right] - \frac{\Pi_2}{\eta_0}y_1(y_1 - H)$$

Using velocity continuity to compute b (again?)   This time we use $u(y_2)|_{top} = u(y_2)|_{mid}$:

$$2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(-\frac{\Pi_2}{K}(y_2-y_2)+\dot{\varepsilon}_0^n\right)^{1/n+1}-\left(-\frac{\Pi_2}{K}(H-y_2)+\dot{\varepsilon}_0^n\right)^{1/n+1}\right]=\frac{\Pi_2}{\eta_0}(y_2^2-y_2H)+b$$

$$2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\dot{\varepsilon}_0^{n+1}-\left(-\frac{\Pi_2}{K}(H-y_2)+\dot{\varepsilon}_0^n\right)^{1/n+1}\right]=\frac{\Pi_2}{\eta_0}(y_2^2-y_2H)+b$$

and since $H-y_2 = H-H/2-\delta = H/2-\delta = y_1$ and

$$y_2^2-y_2H = y_2(y_2-H) = (H/2+\delta)(-y_1) = (-H/2-\delta)y_1 = (-H+H/2-\delta)y_1 = (-H+y_1)y_1$$

so that we indeed recover the same b value as above.

To summarize:

$$
\begin{aligned}
u(y)|_{bot} &= 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(\frac{\Pi_2}{K}(y-y_1)+\dot{\varepsilon}_0^n\right)^{1/n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right] \\[2mm]
u(y)|_{mid} &= \frac{\Pi_2}{\eta_0}(y^2-y) + 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[+\dot{\varepsilon}_0^{n+1} - \left(\dot{\varepsilon}_0^n - \frac{\Pi_2}{K}y_1\right)^{1/n+1}\right] - \frac{\Pi_2}{\eta_0}y_1(y_1-H) \\[2mm]
u(y)|_{top} &= 2\frac{n}{n+1}\frac{K}{\Pi_2}\left[\left(-\frac{\Pi_2}{K}(y+y_2)+\dot{\varepsilon}_0^n\right)^{\frac{1}{n}+1} - \left(-\frac{\Pi_2}{K}(H+y_2)+\dot{\varepsilon}_0^n\right)^{\frac{1}{n}+1}\right] \\[2mm]
\dot{\varepsilon}_{xy}|_{bot} &= \left(\frac{\Pi_2}{K}(y-y_1)+\dot{\varepsilon}_0^n\right)^{1/n} \\[2mm]
\dot{\varepsilon}_{xy}|_{mid} &= \frac{\Pi_2}{\eta_0}\left(y-\frac{H}{2}\right) \\[2mm]
\dot{\varepsilon}_{xy}|_{top} &= -\left(-\frac{\Pi_2}{K}(y-y_2)+\dot{\varepsilon}_0^n\right)^{1/n}
\end{aligned}
$$

Rather interestingly we find that τ_0 does not directly enter the equations above. This can be explained as follows: since we have the relationship

$$
\eta_0 = K\dot{\varepsilon}_0^{n-1} + \frac{\tau_0}{\dot{\varepsilon}_0}
$$

the parameters η_0, $\dot{\varepsilon}_0$, τ_0 and K cannot be all chosen freely. The viscosity η_0 is reached when the strain rate becomes smaller than $\dot{\varepsilon}_0$, so these two parameters have a physical meaning. We set $\eta_0 = 10^{25}$ and $\dot{\varepsilon}_0 = 10^{-17}$. When/if K is zero, then τ_0 can be interpreted as a yield value for a rigid plastic material so we arbitrarily set it to $\tau_0 = 10^7$. Having fixed these parameters we can compute

$$
K = \frac{\eta_0\dot{\varepsilon}_0 - \tau_0}{\dot{\varepsilon}_0^n}
$$

The data used to produce all the following plots is generated by the python program and a gnuplot script to be found in `images/mms/channel_hb/`.

Let's start simple: $n = 1$ In this case the viscosity is given by

$$
\eta_{HB} = \begin{cases} \eta_0 & \dot{\varepsilon}_e \leq \dot{\varepsilon}_0 \\ K + \frac{\tau_0}{\dot{\varepsilon}_e} & \dot{\varepsilon}_e \geq \dot{\varepsilon}_0 \end{cases}
$$

Since $\dot{\varepsilon}_e = \left|\frac{1}{2}\frac{\partial u}{\partial y}\right|$ then

$$
\eta_{HB} = \begin{cases} \eta_0 & \dot{\varepsilon}_e \leq \dot{\varepsilon}_0 \\ K + \frac{2\tau_0}{\left|\frac{\partial u}{\partial y}\right|} & \dot{\varepsilon}_e \geq \dot{\varepsilon}_0 \end{cases}
$$
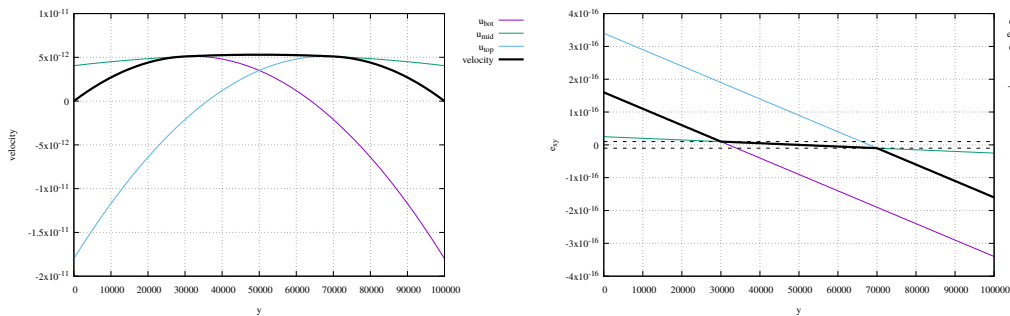
Obtained for $n = 1$ and $\tau_0 = 9e7$. The black lines are the resulting velocity and strain rate profiles obtained by joining the bottom, middle and top functions.

In the following I explore the effect of the τ_0 value (K is calculated correspondingly as we have seen before).
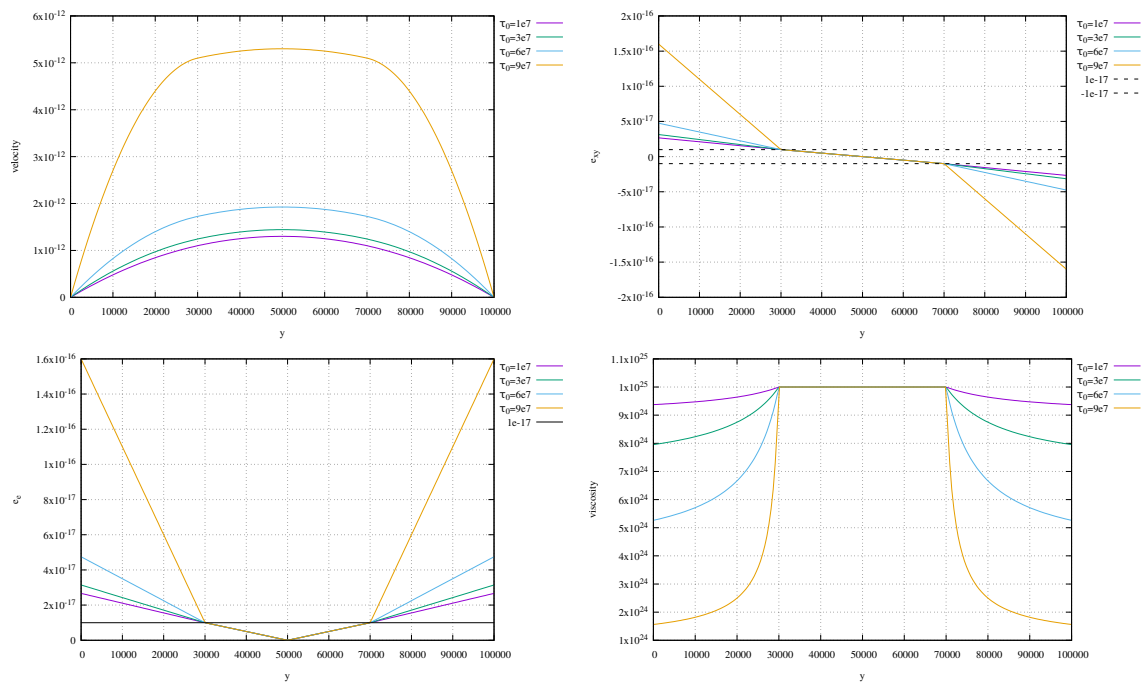
TODO!!!!

Chapter 20

Geodynamical/geophysical benchmarks

20.1 Benchmarks organised by topics

Some published numerical experiments have over time become benchmarks for the entire community. I present in what follows a short list of 'famousx' benchmarks' in the computational geodynamics community.

20.1.1 Subduction

- subduction problems: Spiegelman and Katz [1268] (2006), Schmeling et al. [1200] (2008), van Keken et al. [1409] (2008), Cerpa, Hassani, Gerbault, and Prévost [233] (2014), Glerum, Thieulot, Fraters, Blom, and Spakman [496] (2018), OzBench et al. [1027] (2008), Sime and Wilson [1247] (2020).

- Benchmark of 3D numerical models of subduction against a laboratory experiment: Meriaux *et al.* (2018) [913]

20.1.2 Stokes sphere & sinkers

- the Stokes sphere: Gale manual [787], Aspect manual [48], in visco-plastic fluid: Liu, Muller, and Denn [840] (2002), Besses, Magnin, and Jay [90] (2004). Finite deformation in and around a fluid sphere [1203, 310].

- the sinking block (sinker) Thieulot [1343] (2011), Cerpa, Hassani, Gerbault, and Prévost [233] (2014), Gerya [483] (2010), Gerya and Yuen [481] (2003), May and Moresi [895] (2008), Mishin [927] (2011), Furuichi, May, and Tackley [457] (2011), Maierová [871] (2012), Schuh-Senlis, Thieulot, Cupillard, and Caumon [1221] (2020), Mishin, Vasilyev, and Gerya [928] (2022). (see Section 20.8)

- multiple sinkers [897, 894, 279, 1160]

- Sinking cylinder (2D Stokes sphere): appendix A of [121], [1439].

- hot blob problem [202, 457] (see Section 20.9)

20.1.3 (Visco-)Elastic deformation

- bending of elastic plate/beam [233, 121, 1428, 390, 341, 952, 845]

- flexure of finite length elastic plate [256]

- stress build-up in Maxwell visco-elastic material [478, 256, 390, 341]

- single layer visco-elastic folding [1193, 1428]

- viscous(-elastic) flow around a cylinder in a channel (see Section 20.11)

- Semi-infinite elastic half plane with a circular hole [1419]

- Uniform strip load on elastic material (see Section **??**)

- Jull and McKenzie (1996) [697] parabolic load on viscoelastic half-space (and melt fractions)

- Stress distribution in elastic sphere under equal and opposite loads [1294]

- Love's problem: Becker & Bevis (2004) [69]

- Infinite plate with a circular hole [1487, 1099]

- Square plate with a crack subjected to a horizontal tensile traction [845]

- Hollow sphere under internal pressure, see Section 20.33

20.1.4 Convection

- 2D Rayleigh-Benard convection (see Section 20.28).

- 2D Rayleigh-Benard convection, lateral heating, 30+ codes: Vahl Davis and Jones [1390] (1983).

- 2D Rayleigh-Benard convection with nonlinear rheology: Tosi et al. [1365] (2015), ASPECT manual [48], Trim, Butler, and Spiteri [1372] (2021), STONE 28, Davies, Kramer, Ghelichkhan, and Gibson [325] (2022), Sime and Wilson [1247] (2020), Candioti, Schmalholz, and Duretz [221] (2020).

- 3D convection at infinite Prandtl number with modest viscosity variation: Busse et al. [207] (1994), Trompert and Hansen [1373] (1998), Kameyama, Kageyama, and Sato [706] (2005), O'Neill, Moresi, Müller, Albert, and Dufour [1006] (2006), Kronbichler, Heister, and Bangerth [776] (2012), Trim, Butler, and Spiteri [1372] (2021), Davies, Kramer, Ghelichkhan, and Gibson [325] (2022). This benchmark is carried out in Stone **??**.
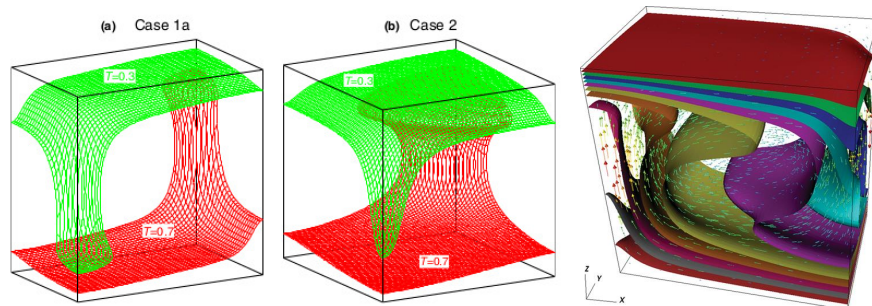
Left: Taken from Kameyama *et al.* (2005). a) Isothermal surfaces obtained for the benchmark calculations of stationary convections in Busse *et al.* .

(1993). (a) Case 1a is for constant viscosity, while (b) Case 2 is for modestly temperature-dependent viscosity whose viscosity contrast is 20. The

calculations were carried out with (a) 64x32x64 and (b) 64x64x64 mesh divisions. Right: Taken from Kronbichler *et al.* (2012).

- Numerical simulations of three-dimensional thermal convection in a fluid with strongly temperature-dependent viscosity: Ogawa *et al.* [1011, 706]

- Convection in 2D-box [467] (see Section 20.7)

- Onset of convection [48]

- mantle convection in 3D spherical shell: Ratcliff, Schubert, and Zebib [1113] (1996), Iwase [661] (1996), Zhong, Zuber, Moresi, and Gurnis [1523] (2000), Yoshida and Kageyama [1488] (2004), Stemmer, Harder, and Hansen [1287] (2006), Choblet, Čadek, Couturier, and Dumoulin [254] (2007), Zhong, McNamara, Tan, Moresi, and Gurnis [1521] (2008), Kameyama, Kageyamab, and Sato [703] (2008), Wright, Flyer, and Yuen [1471] (2010), Davies, Davies, Bollada, Hassan, Morgan, and Nithiarasu [327] (2013), Burstedde et al. [203] (2013), Arrial, Flyer, Wright, and Kellogg [30] (2014), Liu and King [847] (2019), Trim, Butler, and Spiteri [1372] (2021), Euen, Liu, Gassmöller, Heister, and King [407] (2023), Ilangovan, D'Ascoli, Kohl, and Mohr [652] (ildk24).

- Mantle convection with reversing mobile plates [761]

- A comparison of mantle convection models featuring plates [1278]

20.1.5 Free surface & interfaces

- Free surface evolution: Crameri *et al.* (2012) [304], ASPECT manual [48], Schuh-Senlis, Thieulot, Cupillard, and Caumon [1221] (2020).

- relaxation of sinusoidal interface [304, 1157]

20.1.6 Rayleigh-Taylor

- 2D Rayleigh-Taylor convection/instability:
Prosperetti [1084] (1981), Travis et al. [1368] (1990), Weinberg and Schmeling [1445] (1992), Poliakov and Podlachikov [1073] (1992), Ogawa [1010] (1993), Conrad and Molnar [296] (1997), van Keken, King, Schmeling, Christensen, Neumeister, and Doin [1406] (1997), De Smet, van Den Berg, Vlaar, and Yuen [342] (2000), Soboutia, Ghodsb, and Arkani-Hamed [1255] (2001), Babeyko, Sobolev, Trumbull, Oncken, and Lavier [38] (2002), Tackley and King [1313] (2003), Bourgouin, Mühlhaus, Hale, and Arsac [132] (2006), Davies, Davies, Hassan, Morgan, and Nithiarasu [326] (2007), Battaglia, Storti, and D'Elia [59] (2008), Deubelbeiss and Kaus [350] (2008), Quinteros, Ramos, and Jacovkis [1095] (2009), Samuel and Evonuk [1177] (2010), Suckale, Nave, and Hager [1302] (2010), Leng and Zhong [815] (2011), Thieulot [1343] (2011), Mishin [927] (2011), Logg, Mardal, and Wells [852] (2012), Maierová [871] (2012), Fuchs and Schmeling [447] (2013), Vynnytska, Rognes, and Clark [1432] (2013), Choi, Tan, Lavier, and Calo [256] (2013), Robey and Puckett [1150] (2019), Robey [1149] (2019), de Montserrat, Morgan, and Hasenclever [341] (2019), Louis–Napoleon, Gerbault, Bonometti, Thieulot, Martin, and Vanderhaeghe [857] (2020), Schuh-Senlis, Thieulot, Cupillard, and Caumon [1221] (2020), Mishin, Vasilyev, and Gerya [928] (2022), Burcet, Oliveira, Afonso, and Zlotnik [189] (2024), ASPECT manual [48].

- 3D Rayleigh-Taylor instability: Furuichi, Kameyama, and Kageyama [455] (2008), von Tscharner and Schmalholz [1428] (2015).

- Polydiapirism [1445, 48]

20.1.7 Couette & Poiseuille flow

- Couette flow with temperature dependent viscosity [390, 341]

- Couette flow with shear heating [390]

- Couette flow of a power-law fluid, Section 2.4 of Saramito [1188]

- Poiseuille flow: [426, 449, 1316] (see Section 20.3)

- Poiseuille-Couette flow: Fuchs and Schmeling [447] (2013)

- Poiseuille flow of a power-law fluid, Section 2.3 of Saramito [1188]

- Channel flow: Mancktelow [880] (2008)

- channel flow (nonlinear): [481, 871, 441, 483, 390]

- Squeezing flow between moving parallel plates: Gupta and Gupta [542] (1977)

- Flow of a Power Law Fluid Through a Tube - page 87 of Macosko [867]

20.1.8 Thermal problems

- thermal diffusion of half-cooling space (see Section 20.14)

- thermal diffusion of Gaussian distribution (see compgeo notes, elefant manual)

- Heat flow around a cylinder (see Section 20.13)

20.1.9 Lid driven cavity problems

- Lid driven cavity Kawaguti [725] (1961), Chorin [257] (1967), Schneider, Raithby, and Yovanovich [1206] (1978), Fortin and Thomasset [431] (1979), Ghia, Ghia, and Shin [487] (1982), Koseff and Street [768] (1984), Botella and Peyret [123] (1998), Xiu and Karniadakis [1476] (2001), Bruneau and Saad [170] (2006), Erturk [405] (2009), Terrel, Scott, Knepley, Kirby, and Wells [1330] (2012).

- Lid driven cavity with analytical solution (see Section 19.21)

- Lid driven cavity with nonlinear rheology [85, 1305]

20.1.10 Visco-plastic problems

- the 'plastic brick' (See section 20.5):
  Lemiale, Mühlhaus, Moresi, and Stafford [810] (2008), Quinteros, Ramos, and Jacovkis [1095] (2009), Kaus [720] (2010), Mishin [927] (2011), Mühlhaus, Shi, Olsen-Kettle, and Moresi [971] (2011), Maierová [871] (2012), Spiegelman, May, and Wilson [1267] (2016), Kaus et al. [724] (2016), Glerum, Thieulot, Fraters, Blom, and Spakman [496] (2018), Fraters, Bangerth, Thieulot, Glerum, and Spakman [441] (2019), Mishin, Vasilyev, and Gerya [928] (2022).

- indentor, punch problem (see Section 20.10):
  Vilotte, Daignieres, and Madariaga [1421] (1982), Vilotte, Daignieres, Madariaga, and Zienkiewicz [1422] (1984), Vilotte, Madariaga, Daignieres, and Zienkiewicz [1423] (1986), Hubert-Ferrari, King, Manighetti, Armijo, Meyer, and Tapponnier [635] (2003), Fournier, Jolivet, Davy, and Thomas [433] (2004), Thieulot, Fullsack, and Braun [1347] (2008), Gerbault [475] (2012), Glerum, Thieulot, Fraters, Blom, and Spakman [496] (2018).

- numerical sandbox [175, 175, 871, 178, 496]

- fractal networks of shear bands: Poliakov and Herrmann [1074] (1994)

20.1.11 Miscellaneous

- 2D Rayleigh-Benard laminar plumes, comparison of laboratory and numerical modeling : Vatteville, van Keken, Limare, and Davaille [1413] (2009)

- 2D Cartesian flow with extremely temperature-dependent viscosity: Moresi and Solomatov [956] (1995), Trim, Butler, and Spiteri [1372] (2021).

- Thin layer entrainment (see Section 20.30)

- 1D compression [952]

- 2D compressible Stokes flow problem [**lee˙13**, 660, 1319, 816, 742, 848]

- Wannier flow [1441, 1485, 233]

- plastic oedometer test [256]

- Three-dimensional folding of an embedded viscous layer in pure shear [424]

- dam-break problem [939, 46, 834, 802, 617, 17, 523, 607, 59]

- Slope stability for elasto-plastic materials [1099]

- Time-dependent flow in an annulus [467] (see Section 20.6)

- Slab detachment benchmark (see Section 20.16)

- 3D Hollow sphere Stokes flow benchmark:
  Thieulot (2017) [1341], Horbach *et al.* (2020) [624], Kramer *et al.* [774]

- Axisymmetric hollow sphere compressible Stokes flow benchmark:
  Machetel & yuen (1989) [865]

- Annulus benchmark [48], [1065]

- Viscosity grooves benchmark [48]

- Latent heat benchmark [48]

- Layered flow with viscosity contrast [48] (see Section 20.17)

- Brittle thrust wedges benchmark [178, 48]

- Laplace equation on a semi infinite plate (see Section 20.15)

- 2D Stokes flow over cavity: Popov & Makeev (2014) [1079].

- Analytical solution for solitary porosity waves: Connolly and Podladchikov [295] (2015)

- Analytical solution for solitary wave of magma: Dannberg & Heister (2016) [321] and refs therein

- Stokes flow caused by the motion of a rigid sphere close to a viscous interface: Danov, Gurkov, Raszillier, and Durst [323] (1998)

- Deformation caused by a closed vertical volcanic pipe [120]

- Linear Stability Analysis for Thermal Convections in Spherical Shells Yu and Wang [1493] (2019)

- Viscous half-space loading [585]

- Nakiboglu and Lambeck (1982) has cylindrical load on variety of rheologies [980]

- generalized half-plane and half-space Cerruti [1004, 1527]

- Analytical Solutions of Displacements Produced by spherically-shaped Internal Overpressure [476]

- Sagging viscous bridge [1300]

- Deformation around a terminating fault in a viscous medium [51]

20.2 Rayleigh–Taylor wave instability benchmark

In Appendix D of Thieulot [1343] (2011) I wrote: This benchmark is based on the analytical solution by Ramberg [1105] (1968) and has been the subject of a numerical investigation by Deubelbeiss and Kaus [350] (2008). It consists of a two-layer system driven by gravity. No-slip boundary conditions are imposed on the top and the bottom of the box, while free slip are imposed on the sides. Fluid 1 (η_1, ρ_1) of thickness h_1 overlays fluid 2 (η_2, ρ_2) of thickness h_2 (with $h_1 + h_2 = L_y$). An initial sinusoidal disturbance of the interface between these layers is introduced and is characterised by an amplitude Δ and a wavelength $\lambda = L_x/2$.
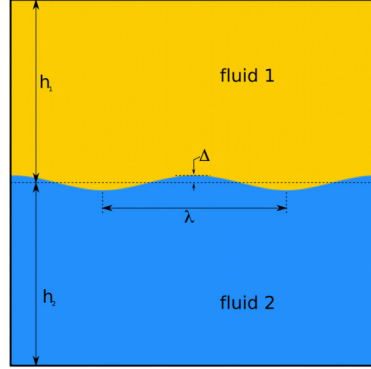
Under these conditions, the velocity of the diapiric growth $\mathbf{v}$ is given by

$$\frac{\mathbf{v}}{\Delta} = -K \frac{\rho_1 - \rho_2}{2\eta_2} h_2 g$$

with K being the dimensionless growth factor, function of ϕ_1, ϕ_2, η_1 and η_2 with

$$K = \frac{-d_{12}}{c_{11}j_{22} - d_{12}i_{21}}$$

and

$$c_{11} = \frac{\eta_1 2\phi_1^2}{\eta_2(\cosh 2\phi_1 - 1 - 2\phi_1^2)} - \frac{2\phi_2^2}{\cosh 2\phi_2 - 1 - 2\phi_2^2} \tag{20.1}$$

$$d_{12} = \frac{\eta_1(\sinh 2\phi_1 - 2\phi_1)}{\eta_2(\cosh 2\phi_1 - 1 - 2\phi_1^2)} + \frac{\sinh 2\phi_2 - 2\phi_2}{\cosh 2\phi_2 - 1 - 2\phi_2^2} \tag{20.2}$$

$$i_{21} = \frac{\eta_1 \phi_2(\sinh 2\phi_1 + 2\phi_1)}{\eta_2(\cosh 2\phi_1 - 1 - 2\phi_1^2)} + \frac{\phi_2(\sinh 2\phi_2 + 2\phi_2)}{\cosh 2\phi_2 - 1 - 2\phi_2^2} \tag{20.3}$$

$$j_{22} = \frac{\eta_1 2\phi_1^2 \phi_2}{\eta_2(\cosh 2\phi_1 - 1 - 2\phi_1^2)} - \frac{2\phi_2^3}{\cosh 2\phi_2 - 1 - 2\phi_2^2} \tag{20.4}$$

$$\phi_1 = \frac{2\pi h_1}{\lambda} \tag{20.5}$$

$$\phi_2 = \frac{2\pi h_2}{\lambda} \tag{20.6}$$

These equations are obtained from Section 20.2 of Gerya [484].

In Thieulot [1343] (2011) this benchmark is ran with $\rho_1 = 3300$, $\rho_2 = 3000$, $\eta_1 = 10^{21}$, η_2 varied between 10^{19} and 10^{27} $L_x = 500km$, L_x varied between 500 and 1500km, leading to $2\pi/3 \leq \phi_1 \leq 2\pi$. A constant resolution of 75×75 elements is chosen and each element contains 5×5 markers [particle in cell technique was used]. This set of measures is conducted for two values of the perturbation amplitude: $\Delta = h_y$ and $\Delta = h_y/3$.
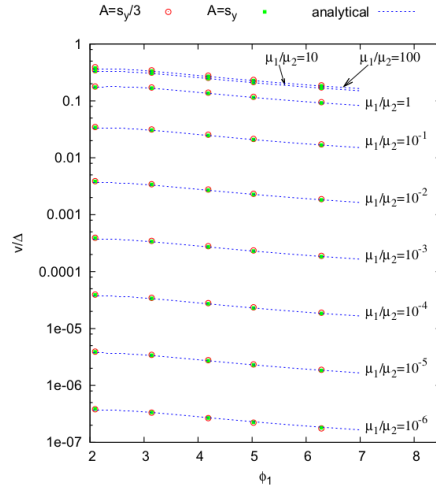
Taken from Thieulot [1343] (2011).

This benchmark is carried out in STONE 40 with slightly different parameters, and also in STONE 120.

20.3 Poiseuille flow

We consider a two-dimensional channel in the x, y plane. The walls are at $y = 0$ and $y = H$ with no-slip boundary conditions. In the absence of gravity, the Stokes equation simplify to

$$-\frac{\partial p}{\partial x} + \frac{\partial}{\partial y}(2\eta_0 \dot{\varepsilon}_{xy}) = 0 \qquad \text{and} \qquad \dot{\varepsilon}_{xy} = \frac{1}{2}\frac{\partial u}{\partial y} \tag{20.7}$$

where we assume the velocity $\vec{v} = (u(y), 0)$. In the case of a Newtonian fluid, the analytical solution is known and the velocity profile is a parabola with zero velocity on the walls and maximum velocity in the middle.

Eq. (20.7) must then be solved

$$\frac{\partial p}{\partial x} = \frac{\partial}{\partial y}\left(2\eta_0 \frac{1}{2}\frac{\partial u}{\partial y}\right) = \eta_0 \frac{\partial^2 u}{\partial y^2} \tag{20.8}$$

We pose $\Pi = \frac{\partial p}{\partial x} < 0$, i.e. there is more pressure applied to the left than to the right of the channel. We then must solve:

$$\frac{\partial^2 u}{\partial y^2} = \frac{\Pi}{\eta_0}$$

The solution is then of the form

$$u(y) = \frac{1}{2}\frac{\Pi}{\eta_0}y^2 + 2ay + b$$

and

$$\dot{\varepsilon}_{xy} = \frac{1}{2}\frac{\Pi}{\eta_0}y + a$$

We will now determine a and b.

The velocity must be zero at $y = 0$ and $y = H$ so

$$u(y = 0) = b = 0$$

and

$$u(y = H) = \frac{1}{2}\frac{\Pi}{\eta_0}H^2 + 2aH = 0$$

or,

$$2a = -\frac{1}{2}\frac{\Pi}{\eta_0}H$$

so

$$\boxed{u(y) = \frac{1}{2}\frac{\Pi}{\eta_0}(y^2 - yH)} \tag{20.9}$$

and

$$\boxed{\dot{\varepsilon}_{xy} = \frac{1}{2}\frac{\Pi}{\eta_0}\left(y - \frac{H}{2}\right)} \tag{20.10}$$

20.4 Relaxation of sinusoidal topography

Following Kramer *et al.* [775, Section 3.1.1] and [1157] the benchmark consists of the relaxation of surface topography in a two-dimensional Cartesian box with an isoviscous fluid. Free slip boundary conditions are imposed on the sides and bottom of the domain. The setup is as follows:
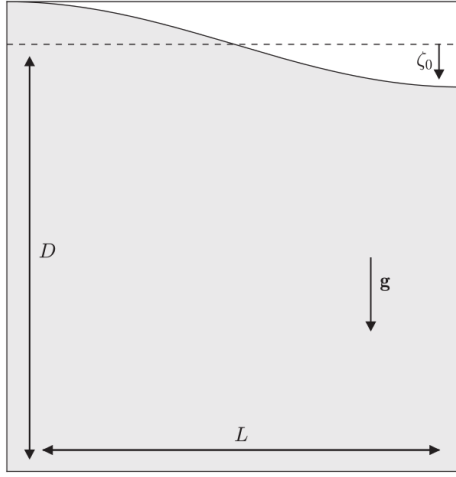
Taken from [1157]. Setup for the free surface relaxation benchmark.
For the tests $\rho = \eta = g = L = D = 1$ and $\xi_0 = 0.005$.
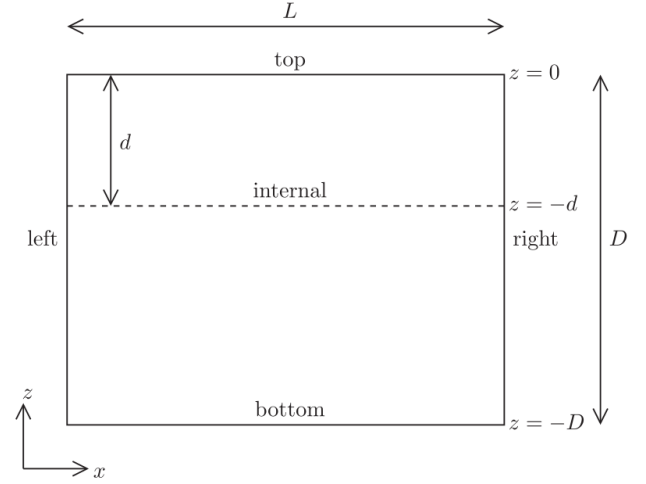
Taken from [775]. $D = 3 \cdot 10^6, \eta = 10^{21}, \rho = 4500, g = 10, \xi_0 = 10^3\,$m, and $L = D/4, D/2, D, 2D, 4D$.

and the infinitesimal sinusoidal perturbations to the free surface is given by

$$\xi(x, t = 0) = \xi_0 \cos\left(\frac{2\pi n x}{L}\right)$$

where n is a wavenumber which is an integer multiple of $1/2$ (taken to be $1/2$ exactly in both cases).

20.5   the plastic brick

📖 Relevant Literature: Hansen [564] (2003), Moresi, Mühlhaus, Lemiale, and May [949] (2007), Lemiale, Mühlhaus, Moresi, and Stafford [810] (2008), Quinteros, Ramos, and Jacovkis [1095] (2009), Kaus [720] (2010), Elesin, Gerya, Artemieva, and Thybo [390] (2010), Mishin [927] (2011), Maierová [871] (2012), Spiegelman, May, and Wilson [1267] (2016), Glerum, Thieulot, Fraters, Blom, and Spakman [496] (2018), Fraters, Bangerth, Thieulot, Glerum, and Spakman [441] (2019), Davies, Kramer, Ghelichkhan, and Gibson [325] (2022), Mishin, Vasilyev, and Gerya [928] (2022), Bangerth et al. [48].

Pretty much all of the brick-type (elasto-)visco-plastic experiments in the literature introduce a weak seed at the bottom of the domain to seed deformation (the shear bands will ultimately stem from it). Dimensioned and dimensionless experiments have been carried out, with or without elastic behaviour, with or without adaptive mesh refinement, with first order and second order quadrilateral elements or Taylor-Hood triangles, with or without Newton algorithm, in extension and compression, with or without time-stepping, with or without viscous lower layer.

Moresi & Mülhaus, 2006 [948]

Compression **Extension**

0.0

0.1

0.3

0.5

tan φ

Moresi *et al.*, 2007 [949]

(A) Newton 20 steps

(B) Picard 360 steps

(C) Picard 120 steps

Plastic strain

0.01 0.03 0.05 0.07 0.09 0.11
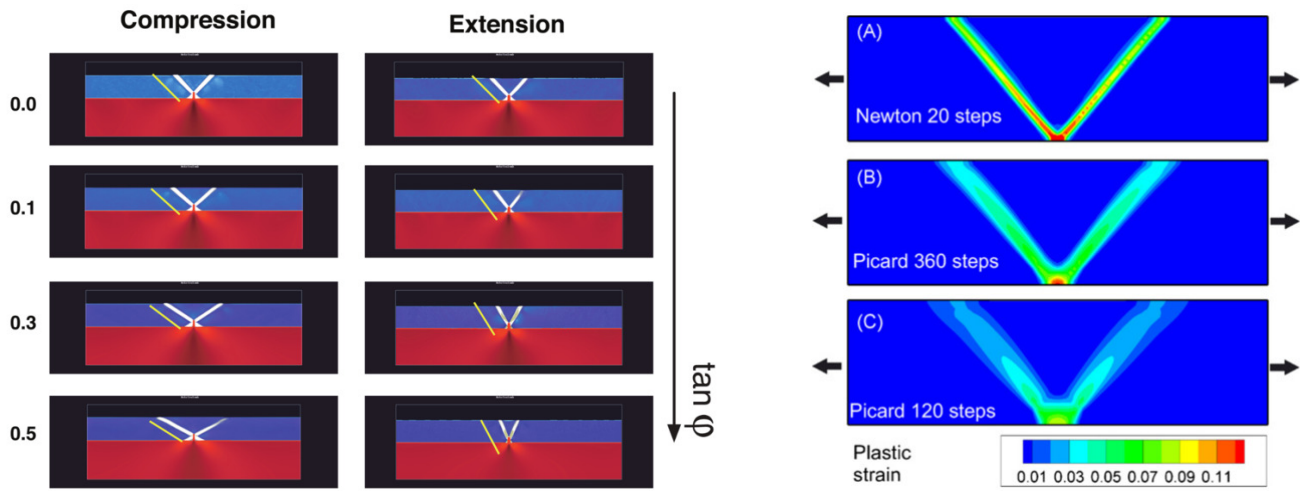
Popov *et al.*, 2008 [1076]
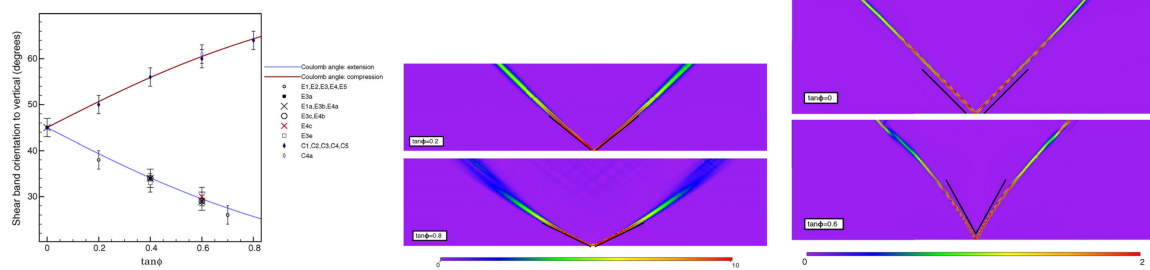
Lemiale *et al.*, 2008 [810]
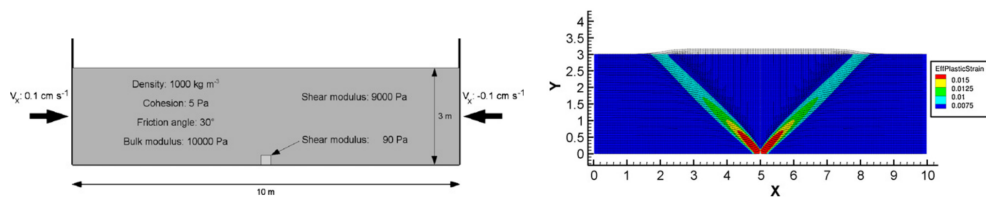
Quinteros *et al.*, 2009 [1095]
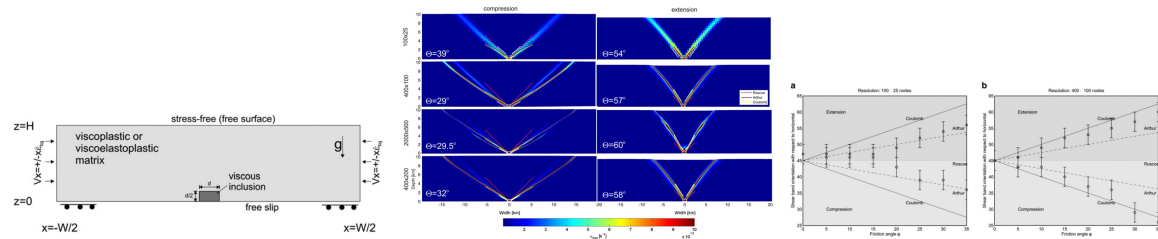
Kaus, 2010 [720]
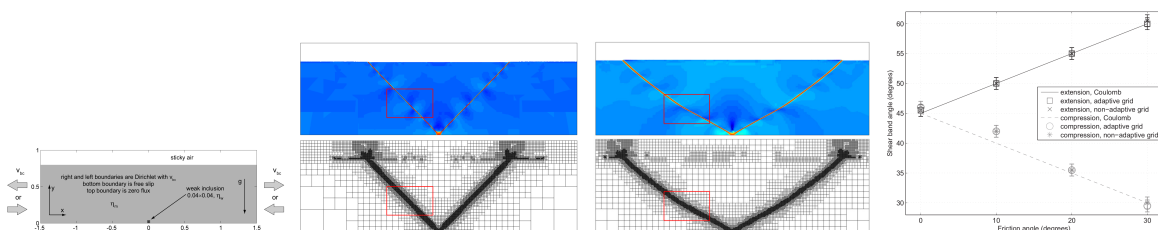
Mishin, phd thesis, 2011 [927]
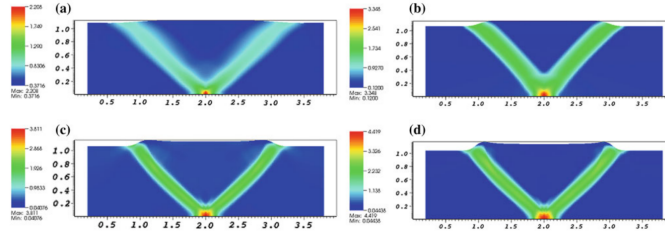
Mühlhaus *et al.* , 2011 [971].
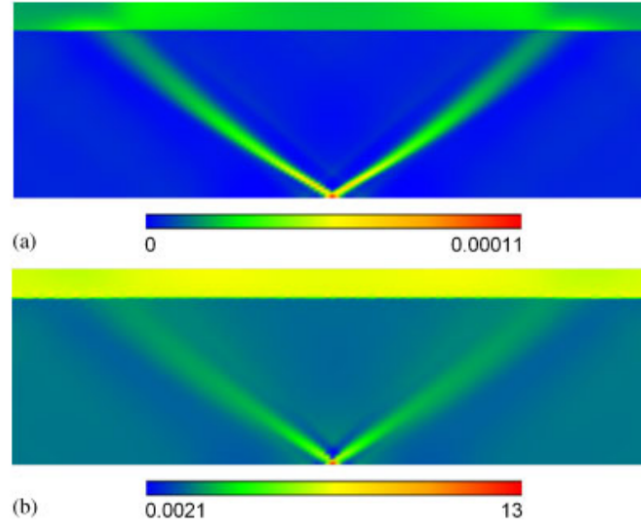
Lemiale *et al.* , 2011 [811].
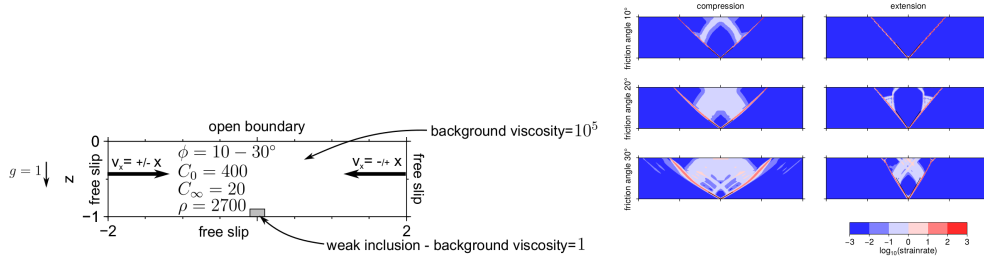
Maierova, phd thesis, 2012 [871]

Mohajeri *et al.* , 2013 [938].

1028

free surface

a)

$\downarrow$ g

viscous inclusion

free slip

b) $\Phi=0°$

45°

c) $\Phi=20°$

compression

35°

d) $\Phi=20°$

extension

55°

1×10^{-17} strain rate 5×10^{-14}

Thieulot, 2014 [1342].

$\boldsymbol{\sigma} \cdot \boldsymbol{n} = \boldsymbol{0}$

η_1

U_0 U_0

H

η_2

$\dot{\epsilon}_{II}$ $(\times10^{-15}\mathrm{s}^{-1})$

0 5 10
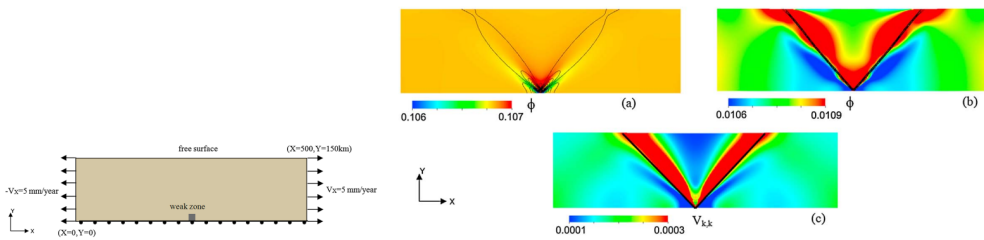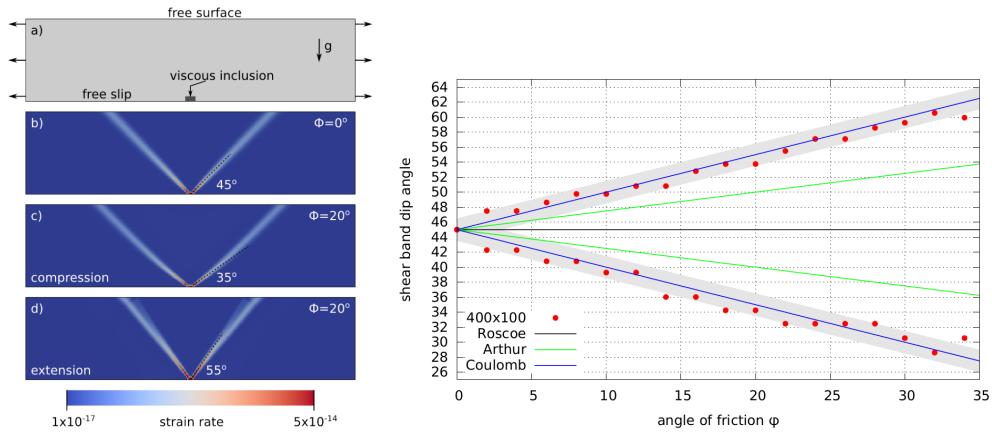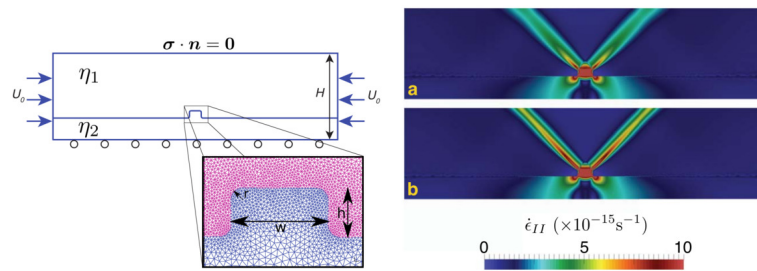
Spiegelman et al., 2016 [1267]

Glerum et al., 2018 [496]

strain-rate [s^{-1}]

velocity [m/s]

viscosity [Pa s]

SPD-factor
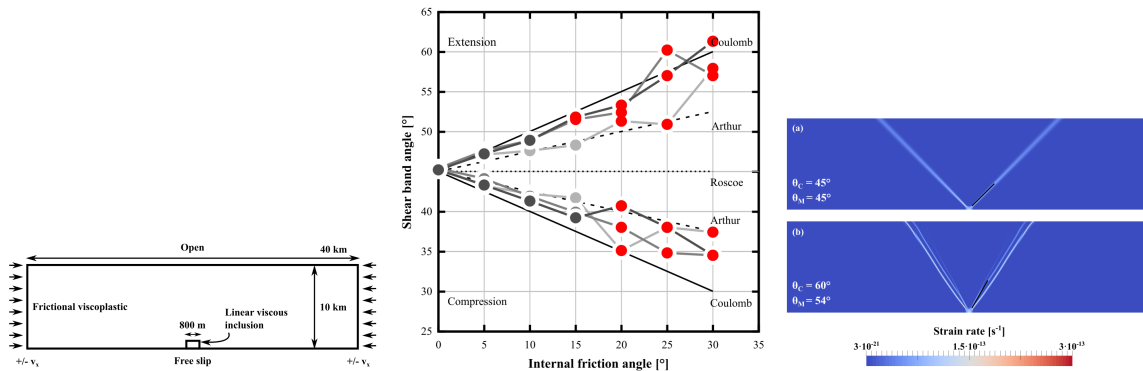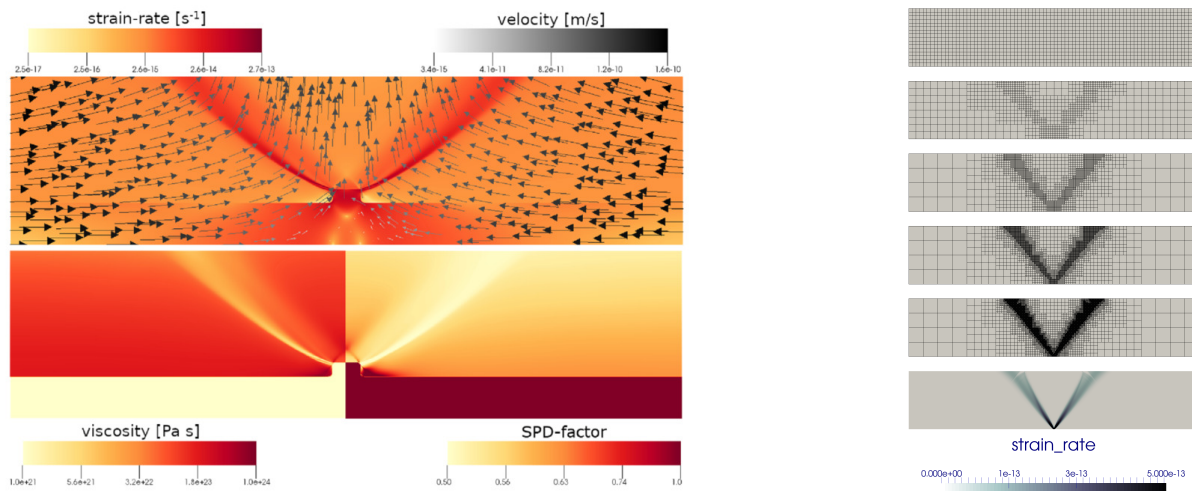
Fraters et al., 2019 [441]

strain_rate

Aspect manual [48]

1029

20.6 Time-dependent benchmark in an annulus

This benchmark is presented in Gassmöller *et al.* [467]. The domain is a 2D annulus with inner and outer radii $R_1 = 1$ and $R_2 = 2$, respectively. In this situation, the incompressible isothermal Stokes equations and their solution can be expressed in a cylindrical coordinate system in terms of the radius r and the azimuthal angle θ. The viscosity is set to $eta = 1$, and the density is given by

$$\rho(r, \theta) = 48r^5 \tag{20.11}$$

The gravity vector is set to

$$\vec{g}(r, \theta) = \frac{r^3}{384}\vec{e}_r + \vec{e}_\theta \tag{20.12}$$

Note that this gravity vector is not the gradient of a gravity potential and consequently not physical. The Stokes system can then be solved using a separation of variables approach and yields

$$\vec{v} = -r^7\vec{e}_\theta \qquad p(r, \theta) = \frac{r^9}{72} - \frac{512}{72} \tag{20.13}$$
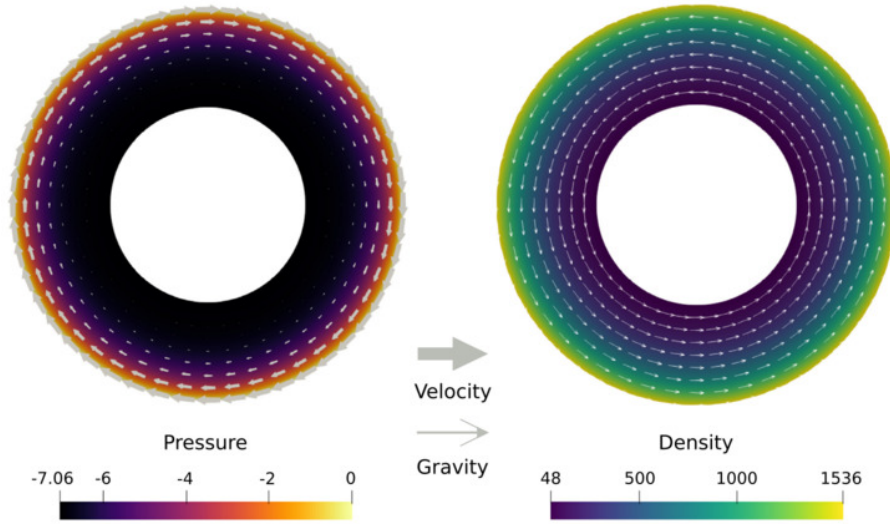
Taken from [467]

Rather importantly, this benchmark was arrived at by means of a stream function (see Section **??**) $\psi(r, \theta) = F(r)G(\theta)$ with $F(r) = r^8/8$ and $G(\theta) = 1$.

20.7 Convection in 2D-box

We start from the following stream function (see Section **??**):

$$\psi(x, y) = \frac{1}{\pi}\sin \pi x \sin \pi y \tag{20.14}$$

which yields:

$$
\begin{aligned}
u(x, y) &= \frac{\partial \psi}{\partial y} = \sin \pi x \cos \pi y \\
v(x, y) &= -\frac{\partial \psi}{\partial x} = -\cos \pi x \sin \pi y
\end{aligned}
\tag{20.15}
$$

1030

The pressure field is
$$p(x, y) = 2\pi \cos(\pi x) \cos(\pi y) \qquad (20.16)$$

with
$$\rho(x, y) = \sin(\pi x)\sin(\pi y) \qquad\qquad g_y = -4\pi^2 \frac{\cos(\pi x)}{\sin(\pi x)} \qquad (20.17)$$

Taken from [467]
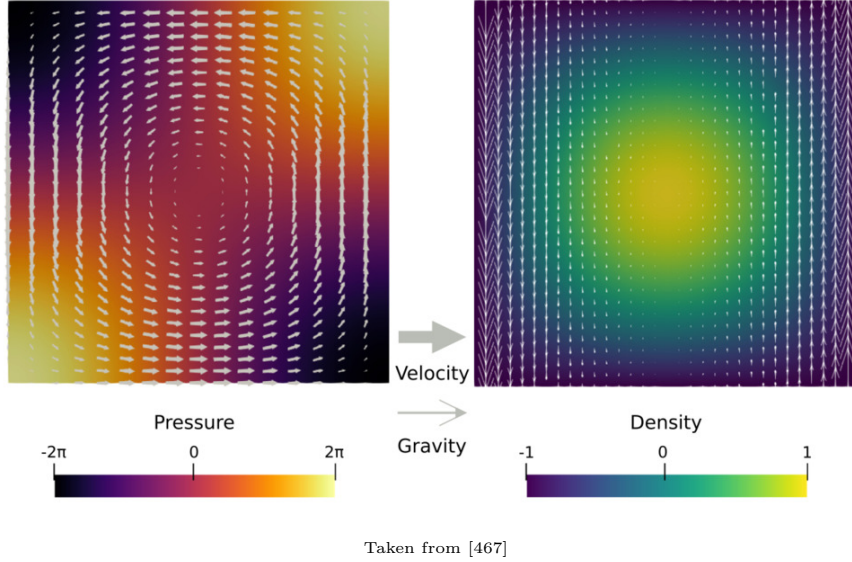
$$
\begin{aligned}
\nu_{rms} &= \sqrt{\frac{1}{L_x L_y} \int_0^1 \int_0^1 (u^2 + v^2)\, dx dy} \\
&= \sqrt{\int_0^1 \int_0^1 (\sin^2(\pi x)\cos^2(\pi y) + \cos^2(\pi x)\sin^2(\pi y))\, dx dy} \\
&= \sqrt{\int_0^1 \sin^2(\pi x)dx \cdot \int_0^1 \cos^2(\pi y)dy + \int_0^1 \cos^2(\pi x)dx \cdot \int \sin^2(\pi y)dy} \\
&= \sqrt{\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}} \\
&= \frac{\sqrt{2}}{2} \\
&\simeq 0.70711... \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (20.18)
\end{aligned}
$$

20.8   The sinker problem

This experiment is not a benchmark stricto sensu since there is no analytical solution. However, it is widely used in the technical literature because of its simple setup and since it allows to test solving strategies. Also, it can conveniently be carried out in both two and three dimensions.

In two dimensions   The time dependent version of the experiment is for instance to be found in Gerya [483] and the same is repeated in Thieulot [1343].

This simple benchmark provides challenging numerical experiments dealing with large viscosity variations within the simulation domain. It consists of a bulk of fluid 1 (η_1, ρ_1) in which a block of fluid 2 (η_2, ρ_2) falls under its own weight. The domain is a square of size $L_x = L_y = 500$km and the

block is initially centred at point $(x = 250\text{km}, y = 400\text{km})$ with size $100 \times 100\text{km}$. Free slip boundary conditions are imposed on all sides of the domain. In [1343] five experiments have been conducted: $\eta_1 = 10^{20}\text{Pa s}, \rho_2 = 3220\text{kg m}^{-3}$; $\eta_1 = 10^{21}\text{Pa s}, \rho_2 = 3300\text{kg m}^{-3}$; $\eta_1 = 10^{22}\text{Pa s}, \rho_2 = 6600\text{kg m}^{-3}$; $\eta_1 = 10^{23}\text{Pa s}, \rho_2 = 3300\text{kg m}^{-3}$; $\eta_1 = 10^{24}\text{Pa s}, \rho_2 = 9900\text{kg m}^{-3}$; while in all experiments the density of the surrounding fluid is $\rho_1 = 3200\text{kg m}^{-3}$ and the viscosity of the block is varied between 10^{19} and $5 \cdot 10^{27}\text{Pa s}$.
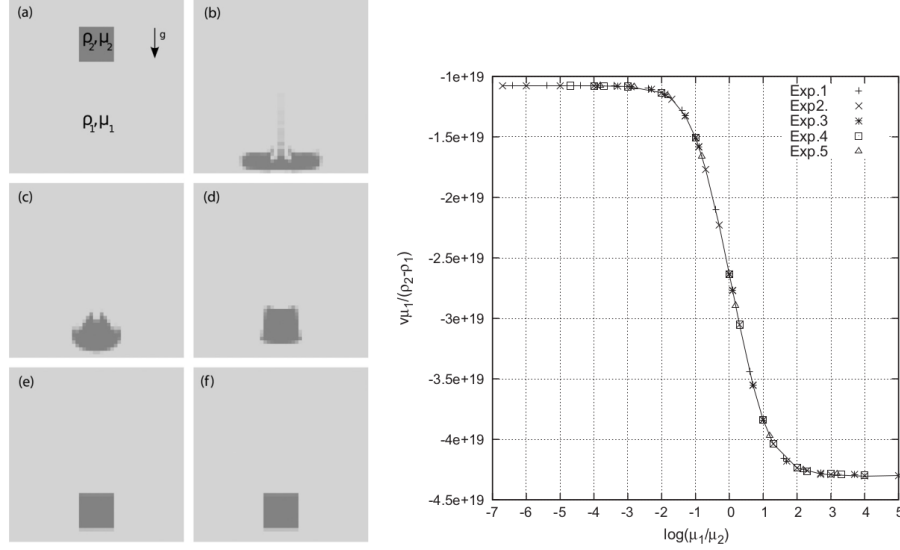
Left: $\eta_1 = 10^{21}\text{Pa s}, \rho_2 = 3300\text{kg m}^{-3}$. (a) Initial setup; (b) $\eta_1 = 10^{21}\text{Pa s}$ at time $t = 10$ Myrs; (c) $\eta_1 = 10^{22}\text{Pa s}$ at time $t = 20$ Myrs; (d) $\eta_1 = 10^{23}\text{Pa s}$ at time $t = 20$ Myrs; (e) $\eta_1 = 10^{25}\text{Pa s}$ at time $t = 20$ Myrs; (f) $\eta_1 = 10^{27}\text{Pa s}$ at time $t = 20$ Myrs. Right: Velocity measurements as a function of the viscosity contrast between surrounding medium and block for all experiments. Taken from [1343]

In three dimensions   Let us look at the sinker experiment from Furuichi *et al.* [457]: The domain is the unit box the origin at the center of the box. A cube with a viscosity $\eta_1 = \Delta\eta$ and density $\rho_1 = 1$ was placed at the middle of the domain defined by $-0.15 \leq x, y, z \leq 0.15$. The material surrounding the cube has the properties $\eta_0 = 1$ and $\rho_0 = 0$. The body force of the momentum equation was taken as $(0, 0, -\rho g)$ with $g = 1$. Along all walls on the domain, free-slip boundary conditions were employed.
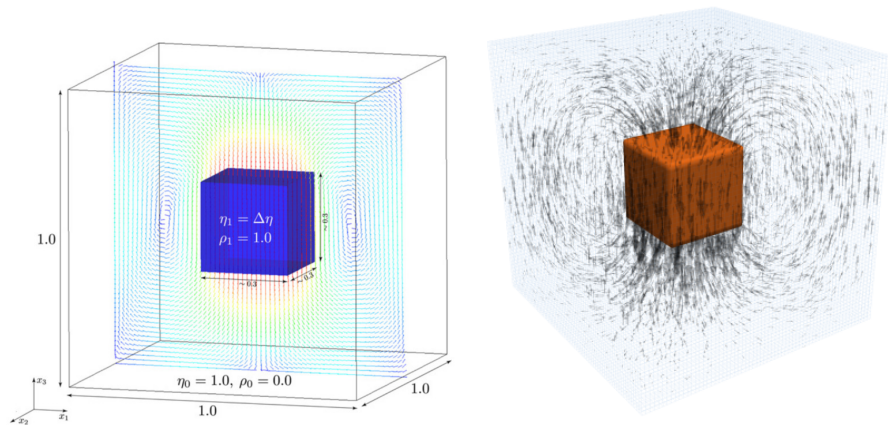
Left: Simulation setup for the 3D falling block (SINKER) problem. The vectors represent computed flow. Taken from Furuichi, May, and Tackley [457] (2011). Right: same experiment in Sanan, May, Mills, et al. [1183] (2022).

Sinking block results for multiple elements   The setup is slightly altered: the domain is 512x512km. The block has size $L_b \times L_b = 128 \times 128\text{km}$, and is centered on $(L_x/2, 3L_y/4)$. Free slip on all sides. Pressure is volume normalised. $|g_y| = 10$. This benchmark is part of ASPECT , and

can therefore be run with $Q_2 \times Q_1$, $Q_2 \times P_{-1}$ and $Q_1 \times P_0$ elements (although the solver does not converge for the latter at high resolutions). Velocity and pressure are measured in the middle of the block (in the case of the $Q_1^+ \times P_0$ element the projected pressure q on the Q_1 is used).

As above, the quantity $|v|\eta_1/\delta\rho$velocity is considered, but this time plotted as a function of the resolution for a fixed η_2 and for various element types. The quantity $p/\delta\rho/L_b$ is also plotted:
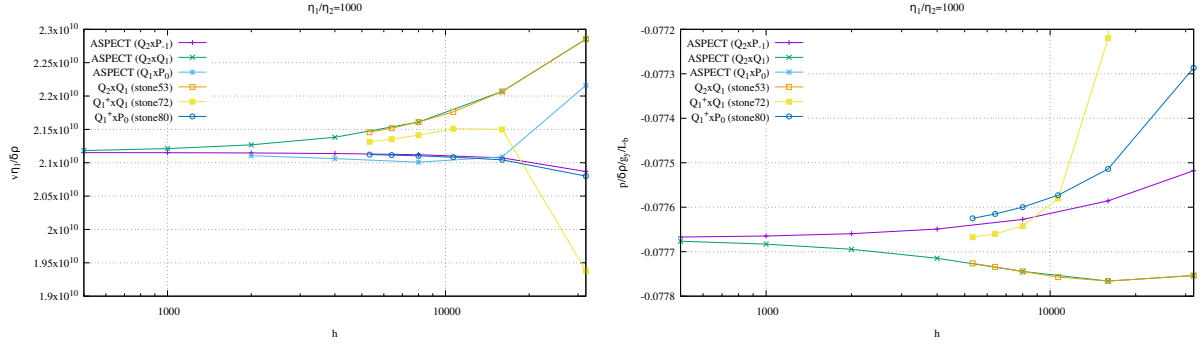
Results for $\rho_1 = 0$, $\rho_2 = \delta\rho = 8$, $\eta_1 = 10^{21}$ and $\eta_2 = 10^{18}$
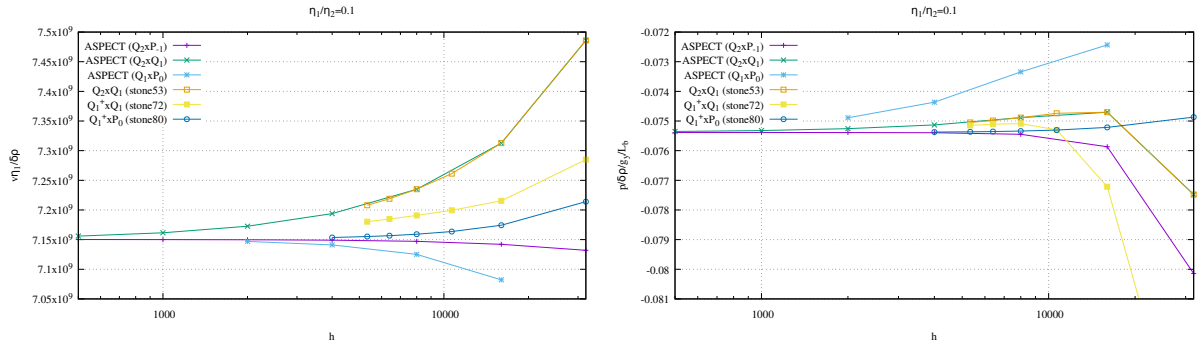
Results for $\rho_1 = 0$, $\rho_2 = \delta\rho = 8$, $\eta_1 = 10^{21}$ and $\eta_2 = 10^{22}$

Results obtained with ASPECT with $\rho_1 = 3200$ and $\rho_2 = \rho_1 + \delta\rho$ are shown here:
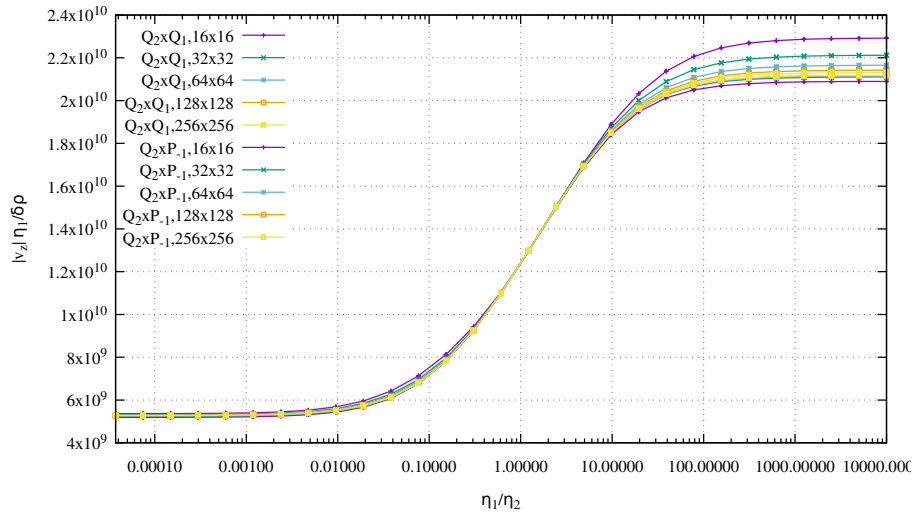
20.9   The hot blob problem

This is a very similar setup as the 3D sinker from the same authors with higher but more diffusive variation of viscosity. The body force is given by $(0, 0, \beta T)$ and where the temperature field T is

defined by $T = \exp(-\gamma(x^2 + y^2 + (z - 0.3)^2))$ with the constant parameters $\beta = 10^6$ and $\gamma = 200$. The temperature-dependent viscosity $\eta = \exp(-\alpha T)$ is employed with the parameter for viscosity contrast α.
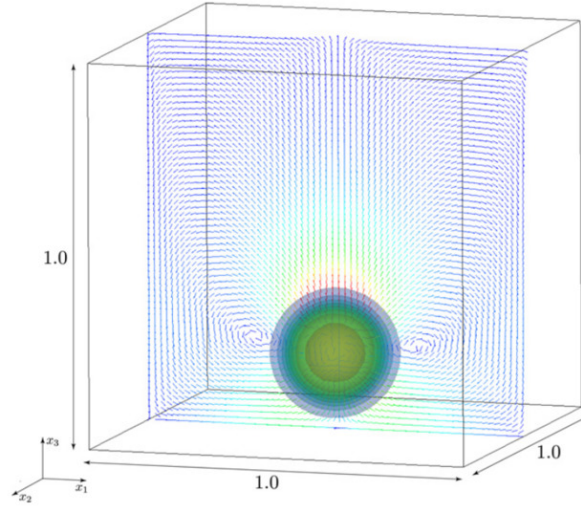
Simulation setting of BLOB problem. Isosurface and vectors represent temperature field
and computed flow respectively. Taken from [457]

20.10 The punch/indentor problem in 2D

The punch benchmark is one of the few boundary value problems involving plastic solids for which there exists an exact solution. Such solutions are usually either for highly simplified geometries (spherical or axial symmetry, for instance) or simplified material models (such as rigid plastic solids) [699].

In this experiment, a rigid punch indents a rigid plastic half space; the slip line field theory gives exact solutions as shown hereunder:
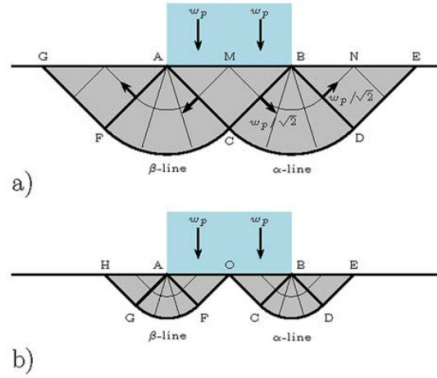
Two-dimensional rigid punch indenting a rigid plastic half space. (a) Prandtl's rigid plastic solution; (b) Hill's solution. Taken from [1347]

The plane strain formulation of the equations and the detailed solution to the problem were derived in the Appendix of [1347] and are also presented in [477] and in [133, Chapt.6]. The two dimensional punch problem has been extensively studied numerically for the past 40 years [1537, 1080, 1536, 1539, 271, 270, 644, 1494, 174, 1097, 496] and has been used to draw a parallel with the tectonics of eastern China in the context of the India-Eurasia collision [1323, 941, 397] or the European Alps [1121]. It is also worth noting that it has been carried out in one form or another in series of analogue modelling articles concerning the same region, with a rigid indenter colliding with a rheologically stratified lithosphere [1322, 1051, 339, 694].
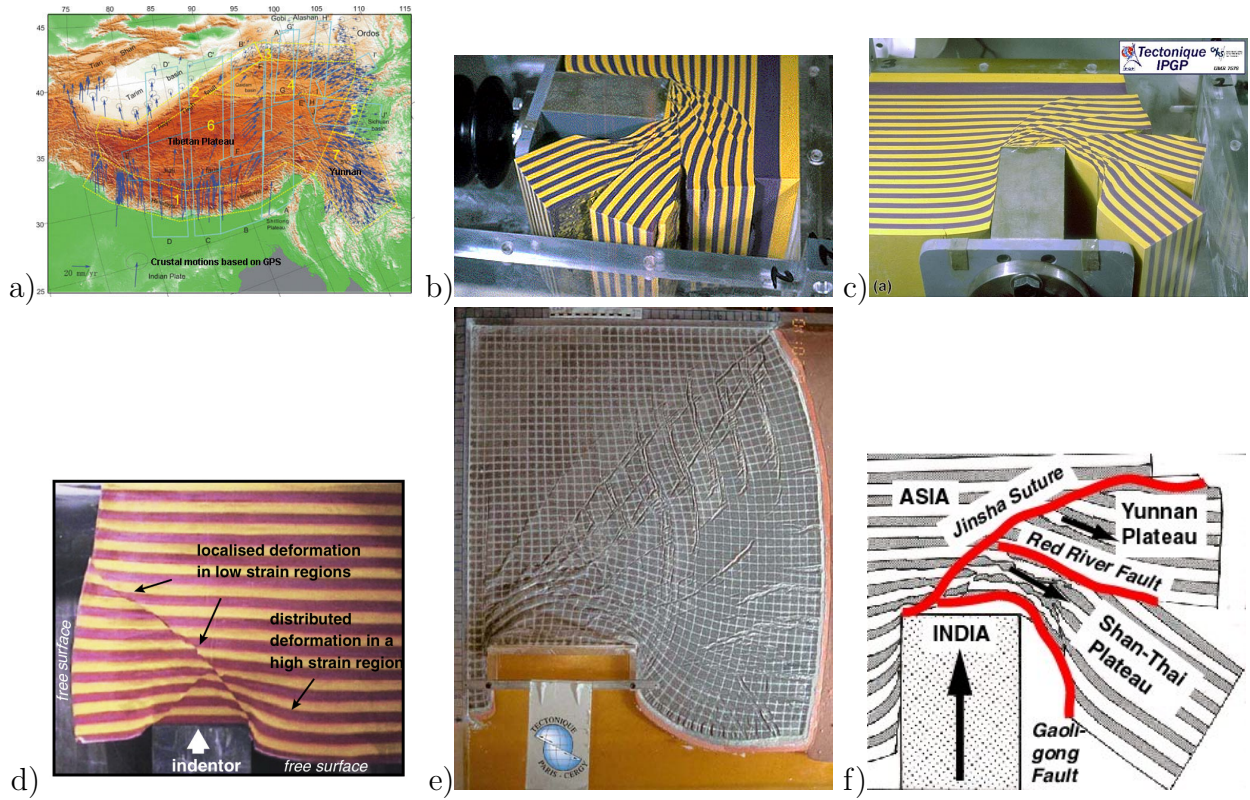
a) b) c) d) e) f)

Numerically, the one-time step punch experiment is performed on a two-dimensional domain of purely plastic von Mises material. Given that the von Mises rheology yield criterion does not depend on pressure, the density of the material and/or the gravity vector is set to zero. Sides are set to free slip boundary conditions, the bottom to no slip, while a vertical velocity $(0, -v_p)$ is prescribed at the top boundary for nodes whose x coordinate is within $[L_x/2 - \delta/2, L_x/2 + \delta/2]$.

The analytical solution predicts that the angle of the shear bands stemming from the sides of the punch is $\pi/4$, that the pressure right under the punch is $1 + \pi$, and that the velocity of the rigid blocks on each side of the punch is $v_p/\sqrt{2}$ (this is simply explained by invoking conservation of mass).
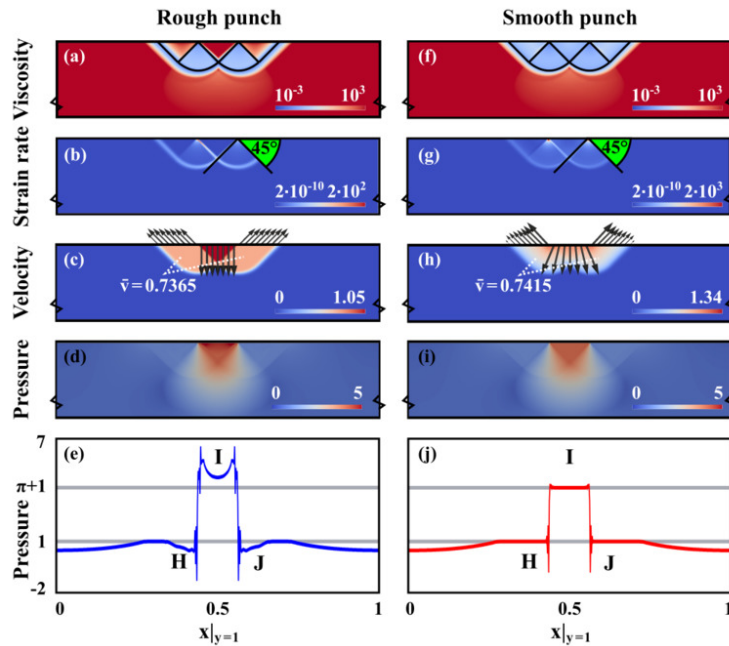
The punch benchmark results after 500 nonlinear iterations for a rough punch (left column) and a smooth punch (right column). (a,f) Viscosity field with analytical slip lines. (b,g) Strain rate norm $\dot{\varepsilon}_e$ with measured shear band angles. (c,h) Velocity magnitude with velocity vectors along the surface of the

1035

domain. (d,i) Pressure field. (e,j) Pressure along the surface of the domain (colored line) and analytical solution values $\pi + 1$ and 1 (grey lines). Taken from

[496]

Remark. *This benchmark is often mentioned or used in the context of bearing capacity, footings, limit state design/analysis [919, 1528, 506, 507, 803, 1239].*

20.11   Viscous flow around a cylinder in 2D and 3D

There are many variants of this problem: 2D in Turek [1382], 3D in John [689]. Many studies focus on Navier-Stokes flow since the cylinder generates vortices at high Reynolds numbers. Steady state solutions at low Re are shown here[1]. Note the interesting benchmark for 2D visco-elastic flow in Beuchert & Podlachikov [92].
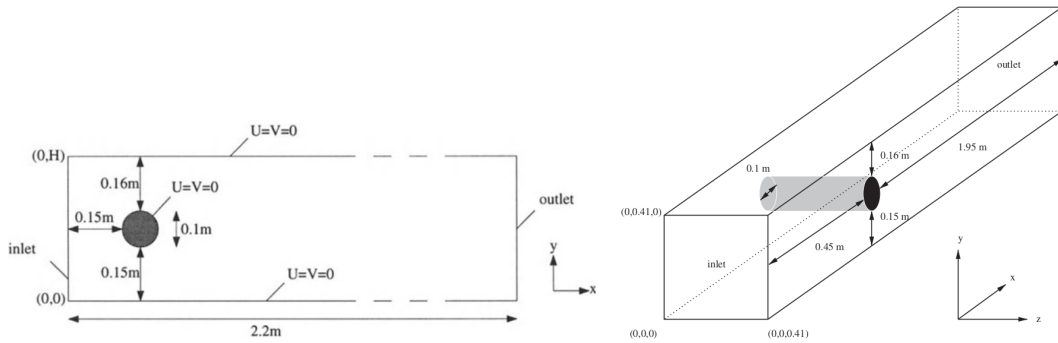
Left: taken from Turek [1382]; Right: taken from John [689]

📙 Relevant Literature: Tachibana & Iemoto (1987) [1310] Schäfer & Turek (1996) [1191]

20.12   Viscous flow around many cylinders in 2D

20.13   Heat flow around a cylinder

The domain is a 2D Cartesian box of size 8x4. The Stokes equations are not solved and the following velocity is prescribed:

$$u(x,y) = U_\infty \left(1 - \frac{x^2 - y^2}{(x^2 + y^2)^2}\right) \tag{20.19}$$

$$v(x,y) = -2U_\infty \frac{xy}{(x^2 + y^2)^2} \tag{20.20}$$

Note that the coordinates in the equation above are actually centered on the middle of the cylinder (this also means that we never divide by zero outside the cylinder!)

Boundary conditions are as follows: $T = 0$ is imposed at the top and bottom of the domain. $T = 1$ is imposed inside a disc centered at (2,2) with radius 1. Further: $k = 0.01$, $C_p = 1$, $\rho = 1$, CFL number is 0.1.

---

[1]upofthetestanddatameasurement
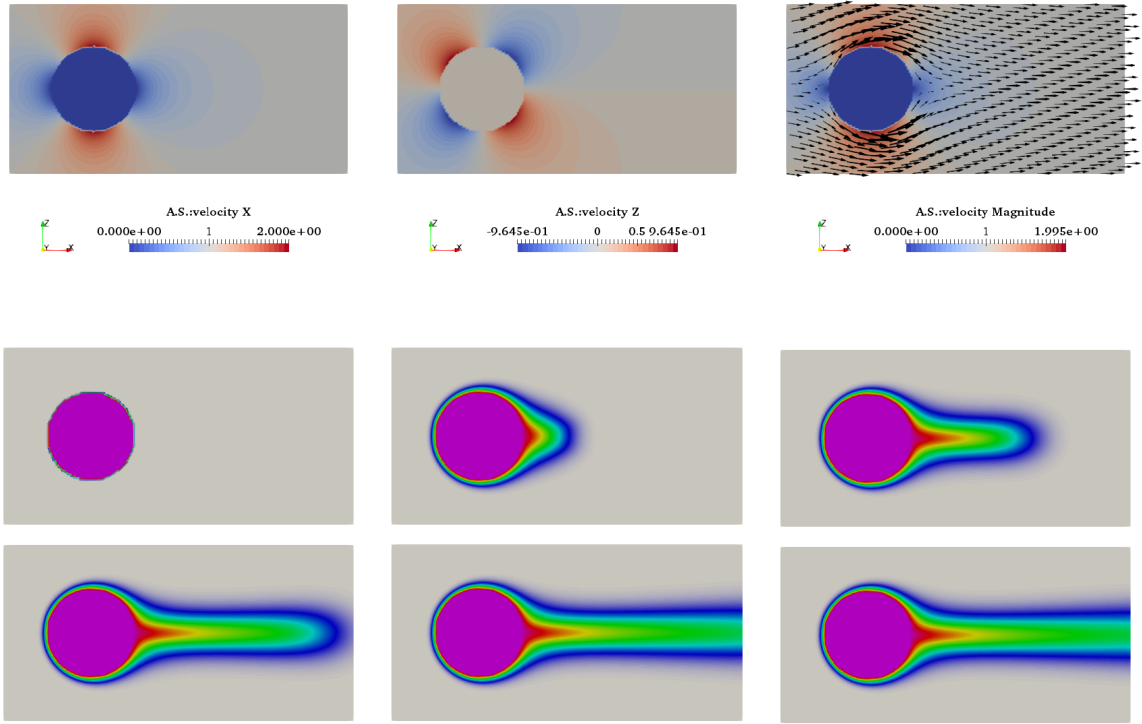
Time evolution of the temperature field. Results obtained with ELEFANT (unpublished)

This is carried out in STONE 65.

20.14 Thermal diffusion of half-cooling space

This is a simple 1D experiment which solution is (for instance) available in Turcotte & Schubert [1378] and is also presented in Choi *et al.* [256].

The domain is 100km deep. $T_0 = 0°$C is prescribed a the surface and $T_m = 1300°$C is prescribed at the bottom. The initial temperature is $T(y) = 1300°$C. The material is characterised by $\rho = 1000$kg/m^3, $C_p = 1000$J/kg/K, $k = 1$J/m/K. The time-dependent solution is given by:

$$T(y,t) = T_0 + (T_0 - T_m)\text{erf}\left(\frac{y}{2\sqrt{kt/\rho C_p}}\right) \tag{20.21}$$
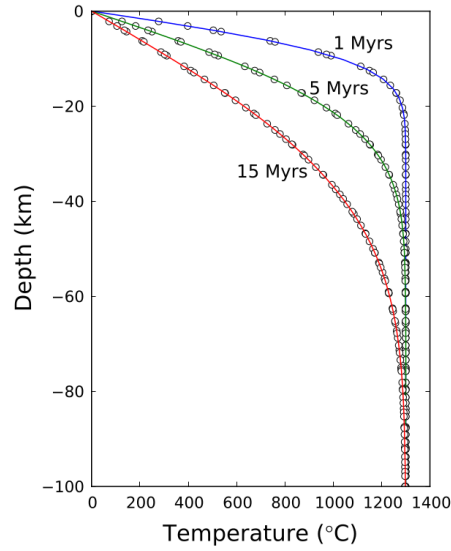
Thermal diffusion of half space cooling plate. The temperature profiles in the analytical solution at 1, 5, and 15 Myrs are plotted in solid lines. The results from DynEarthSol2D are plotted in circles. Taken from [256]

20.15   Laplace equation on a semi infinite plate

This experiment is based on a 2nd year mathematics lecture I give at Utrecht University. One wishes to solve the Laplace equation for temperature on the following plate subject to the indicated boundary conditions:
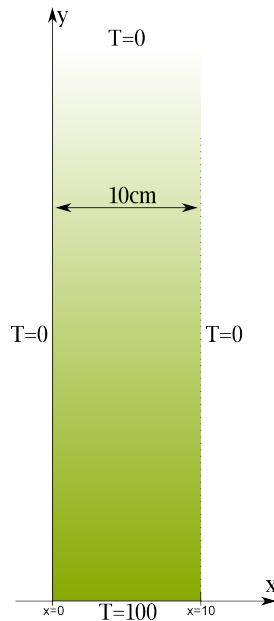
The temperature satisfies the 2D Laplace equation inside the plate:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \tag{20.22}$$

We could try to solve the equation by using a tentative solution of the form:

$$T(x, y) = \theta(x)\Phi(y) \tag{20.23}$$

⚠ We do not *know* the solution is of this form.

We substitute (2) into (1) and obtain:

$$\Phi \frac{\partial^2 \theta}{\partial x^2} + \theta \frac{\partial^2 \Phi}{\partial y^2} = 0$$

Dividing by $\theta\Phi$ gives:

$$\frac{1}{\theta} \frac{\partial^2 \theta}{\partial x^2} + \frac{1}{\Phi} \frac{\partial^2 \Phi}{\partial y^2} = 0$$

Separation of variables: we say that each term is a constant because the first term is a function of x only and the second a function of y only. We then write

$$\frac{1}{\theta} \frac{\partial^2 \theta}{\partial x^2} = -\frac{1}{\Phi} \frac{\partial^2 \Phi}{\partial y^2} = -k^2$$

where k is called the separation constant. This leads to

$$\frac{\partial^2 \theta}{\partial x^2} + k^2 \theta = 0$$

$$\frac{\partial^2 \Phi}{\partial y^2} - k^2 \Phi = 0$$

- The solution to the first one is $\theta(x) = \sin kx$ or $\theta(x) = \cos kx$

- The solution to the second one is $\Phi(x) = e^{kx}$ or $\Phi(x) = e^{-kx}$

The general solution writes:

$$T(x,y) = \theta(x)\Phi(y) = \left\{ \begin{array}{c} \sin kx \\ \cos kx \end{array} \right\} \left\{ \begin{array}{c} e^{ky} \\ e^{-ky} \end{array} \right\}$$

We can now use the b.c. to find the solution to the Laplace equation.

- Since $T \to 0$ when $y \to \infty$ then e^{ky} unacceptable.

- Since $T = 0$ when $x = 0$ then $\cos kx$ unacceptable.

so

$$T(x,y) = \sin(kx)\, e^{-kx}$$

We finally use $T = 0$ at $x = 10$ which leads to $10k = n\pi$, i.e.:

$$T(x,y) = \sin(\frac{n\pi x}{10})\, e^{-n\pi y/10}$$

⚠ Problem: the solution does not satisfy $T(x,0) = 100$. However, a linear combination of solutions is still a solution ! Let's find such a combination which satisfies the b.c. at $y = 0$:

$$T(x,y) = \sum_{n=1}^{\infty} b_n \sin(\frac{n\pi x}{10})\, e^{-n\pi y/10}$$

We impose then $T(x,0) = 100$:

$$100 = \sum_{n=1}^{\infty} b_n \sin(\frac{n\pi x}{10})$$

This is the Fourier sine series of $f(x) = 100$ with $l = 10$ (chapter 7.9 of Boas).

The coefficient b_n is then given by

$$b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx = \frac{2}{10} \int_0^l 100 \sin \frac{n\pi x}{10} dx = \begin{cases} 400/n\pi & \text{odd n} \\ 0 & \text{even n} \end{cases}$$

Finally (!):

$$T(x,y) = \frac{400}{\pi} \left(e^{-\pi y/10} \sin(\frac{\pi x}{10}) + \frac{1}{3} \sin(\frac{3\pi x}{10}) e^{-3\pi y/10} + \ldots \right)$$

The simulation has been run with a 10x50 domain. All coefficients of the temperature equation are set to 1, and the Stokes equation is not solved. The timestep is fixed to $dt = 0.1$. Resolution is 32x160.
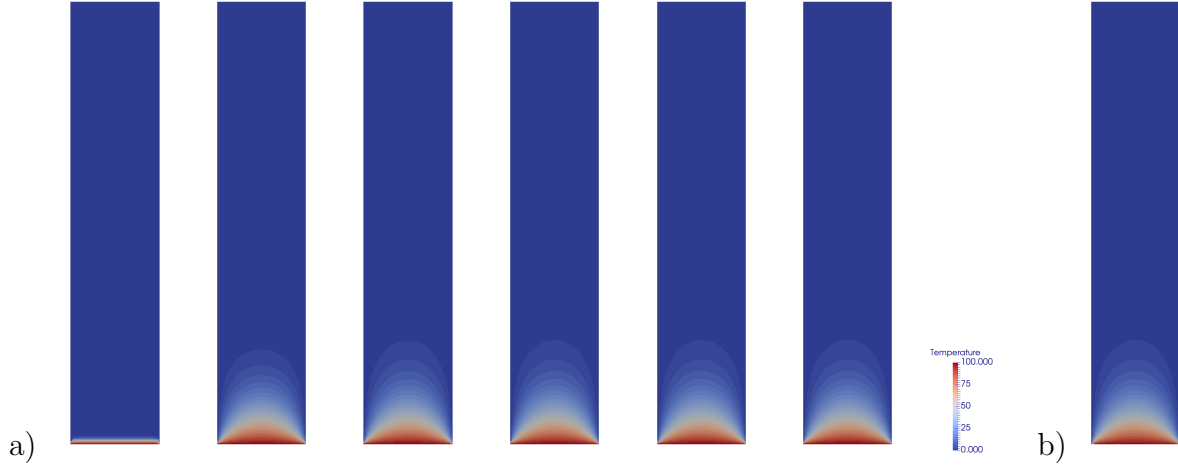
a) time evolution of the temperature field; b) analytical steady state solution

20.16   Slab detachment benchmark

🔹 Relevant Literature: Schmalholz [1192] (2011), Glerum, Thieulot, Fraters, Blom, and Spakman [496] (2018), Lee and So [805] (2019), ASPECT manual [48], STONE 26.
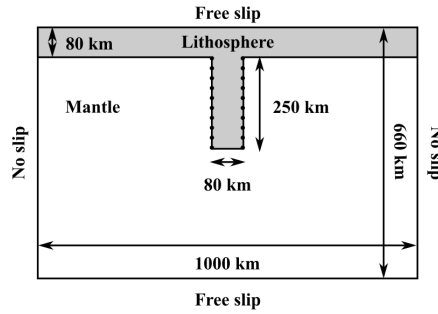
The detachment benchmark model setup of Schmalholz [1192]: a symmetric system of nonlinear viscous lithosphere with a vertical slab extending into a linear viscous mantle. The top and bottom boundaries are free slip, while the vertical boundaries are no slip. Taken from [496].

20.17   Layered flow with viscosity contrast

The idea behind this benchmark is to construct an analytical solution to the incompressible Stokes equation in the case where the viscosity field showcases a viscosity contrast at location $y = y_0$ whose amplitude and width can be controlled. The viscosity is defined as

$$\eta(y) = \frac{1}{\frac{1}{\pi} \tan^{-1}(\frac{y-y_0}{\beta}) + 1/2 + \epsilon}$$

where β and ϵ are parameters.
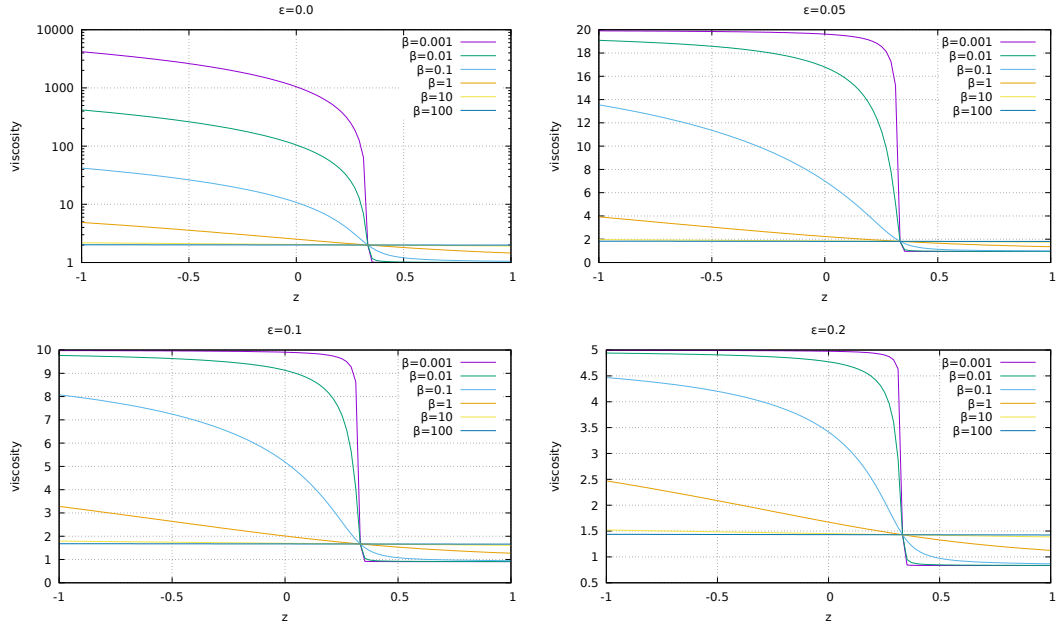
Viscosity profiles for different values of β and ϵ for $y_0 = 1/3$. When β is very large, the viscosity essentially converges to $\sim (1/2 + \epsilon)^{-1}$. β controls the width of the transition while ϵ controls the amplitude of the viscosity variation.

The flow is assumed to take place in an infinitely long pipe (in the horizontal direction) and bound by $y = -1$ and $y + 1$. At the bottom we impose $v_x(y = -1) = 0$ while we impose $v_x(y = +1) = 1$ at the top. The density is set to 1 while the gravity is set to zero. Under these assumptions, the flow velocity and pressure fields are given by:

$$
\begin{aligned}
v_x(x, y) &= \frac{1}{2\pi}\left(-\beta C_1 \log[\beta^2 + (z - y_0)^2] + 2(z - y_0)C_1 \tan^{-1}\frac{z - y_0}{\beta} + \pi(1 + 2\epsilon)zC_1 + C_2\right)\\
v_y(x, y) &= 0\\
p(x, y) &= 0
\end{aligned}
$$

(20.24)

where C_1 and C_2 are integration constants:

$$
C_1 = 2\pi\left[\beta \log[\beta^2 + (1 + y_0)^2] - 2(1 + y_0)\tan^{-1}\frac{1 + y_0}{\beta} - \beta \log[\beta^2 + (1 - y_0)^2] + 2(1 - y_0)\tan^{-1}\frac{1 - y_0}{\beta} + 2\right]
$$

$$
C_2 = \left[\beta \log[\beta^2 + (1 + y_0)^2] - 2(1 + y_0)\tan^{-1}\frac{1 + y_0}{\beta} + \pi(1 + 2\epsilon)\right]C_1,
$$
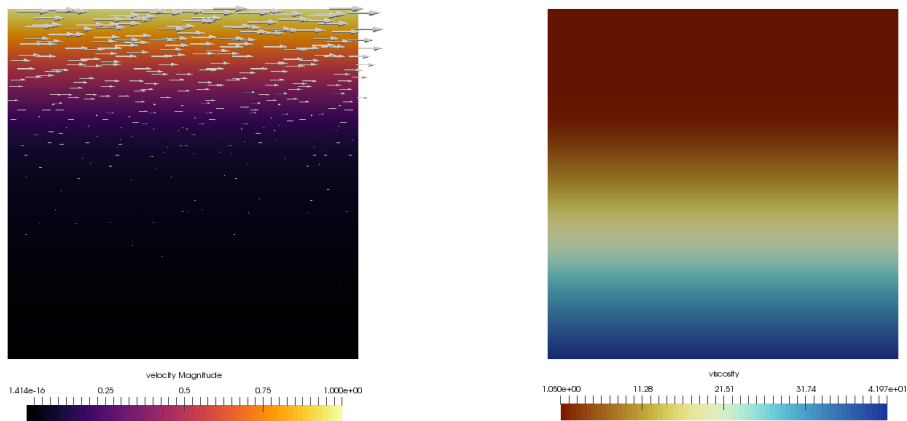
Velocity and viscosity fields

1041

Analytical derivations The flow takes place in the horizontal direction and is infinite in the this direction too so that:

$$\vec{v} = (u(y), 0)$$

The strain rate tensor is then given by:

$$\dot{\varepsilon} = \frac{1}{2} \begin{pmatrix} 0 & du/dy \\ du/dy & 0 \end{pmatrix}$$

The momentum equation then becomes:

$$\vec{\nabla} \cdot (2\eta\dot{\varepsilon}) - \vec{\nabla}p = \vec{\nabla} \cdot \left[\eta(y) \begin{pmatrix} 0 & du/dy \\ du/dy & 0 \end{pmatrix} \right] - \vec{\nabla}p = \rho\vec{g}$$

On the vertical axis, when the gravity is zero, the equation is automatically verified when the pressure is zero. On the horizontal axis:

$$\frac{d}{dy}\left(\eta(y)\frac{du}{dy} \right) = 0$$

Then

$$\eta(y)\frac{du}{dy} = C_1$$

or,

$$\frac{du}{dy} = \frac{C_1}{\eta(y)} = C_1 \left(\frac{1}{\pi} \tan^{-1} \frac{y - y_0}{\beta} + 1/2 + \epsilon \right)$$

so that the velocity is given by:

$$u(y) = \frac{1}{\pi}(y \tan^{-1}((y - y_0)/\beta) - y_0 \tan^{-1}((y - y_0)/\beta) - 0.5 * \beta \log(\beta^2 + y^2 - 2yy_0 + y_0^2) + \pi y(\epsilon + 0.5))$$

$$u(z) = \frac{1}{2\pi}\left(-\beta C_1 \log[\beta^2 + (z - y_0)^2] + 2(z - y_0)C_1 \tan^{-1}\frac{z - y_0}{\beta} + \pi(1 + 2\epsilon)zC_1 + C_2 \right)$$

where C_1 and C_2 are integration constants. I wish to impose $u(z = -1) = 0$ and $u(z = +1) = 1$:

$$\frac{1}{2\pi}\left(-\beta C_1 \log[\beta^2 + (-1 - y_0)^2] + 2(-1 - y_0)C_1 \tan^{-1}\frac{-1 - y_0}{\beta} - \pi(1 + 2\epsilon)C_1 + C_2 \right) = 0$$

$$\frac{1}{2\pi}\left(-\beta C_1 \log[\beta^2 + (1 - y_0)^2] + 2(1 - y_0)C_1 \tan^{-1}\frac{1 - y_0}{\beta} + \pi(1 + 2\epsilon)C_1 + C_2 \right) = 1$$

or,

$$-\beta C_1 \log[\beta^2 + (-1 - y_0)^2] + 2(-1 - y_0)C_1 \tan^{-1}\frac{-1 - y_0}{\beta} - \pi(1 + 2\epsilon)C_1 + C_2 = 0$$

$$-\beta C_1 \log[\beta^2 + (1 - y_0)^2] + 2(1 - y_0)C_1 \tan^{-1}\frac{1 - y_0}{\beta} + \pi(1 + 2\epsilon)C_1 + C_2 = 2\pi$$

or,

$$-\beta C_1 \log[\beta^2 + (-1 - y_0)^2] + 2(1 + y_0)C_1 \tan^{-1}\frac{1 + y_0}{\beta} - \pi(1 + 2\epsilon)C_1 + C_2 = 0$$

$$-\beta C_1 \log[\beta^2 + (1 - y_0)^2] + 2(1 - y_0)C_1 \tan^{-1}\frac{1 - y_0}{\beta} + \pi(1 + 2\epsilon)C_1 + C_2 = 2\pi$$

or,

$$-\beta C_1 \log(\beta^2 + (1 + y_0)^2) + 2(1 + y_0)C_1 \tan^{-1}((1 + y_0)/\beta) - \pi(1 + 2\epsilon)C_1 + C_2 = 0$$

$$-\beta C_1 \log(\beta^2 + (1 - y_0)^2) + 2(1 - y_0)C_1 \tan^{-1}((1 - y_0)/\beta) + \pi(1 + 2\epsilon)C_1 + C_2 = 2\pi$$

I can now substract the first line from the second line:

$$\beta C_1 \log(\beta^2+(1+y_0)^2)-2(1+y_0)C_1 \tan^{-1}((1+y_0)/\beta)-\beta C_1 \log(\beta^2+(1-y_0)^2)+2(1-y_0)C_1 \tan^{-1}((1-y_0)/\beta)+2\pi(1$$

i.e.,

$$C_1 = 2\pi \left[\beta \log[\beta^2 + (1 + y_0)^2] - 2(1 + y_0) \tan^{-1}[\frac{1 + y_0}{\beta}] - \beta \log[\beta^2 + (1 - y_0)^2] + 2(1 - y_0) \tan^{-1}[\frac{1 - y_0}{\beta}] + 2\pi \right.$$

and then

$$C_2 = \beta C_1 \log(\beta^2 + (1 + y_0)^2) - 2(1 + y_0)C_1 \tan^{-1}((1 + y_0)/\beta) + \pi(1 + 2\epsilon)C_1$$

20.18 The annulus convection benchmark # 1

We wish to solve the Stokes equation in an annulus of inner radius R_1 and outer radius R_2 with the following boundary conditions:

- Inner boundary: $\mathbf{v}_r(R_1, \theta) = 0$

- Outer boundary: $\mathbf{v}_r(R_2, \theta) = 0$

We then postulate

$$\mathbf{v}_\theta(r, \theta) = f(r)\cos(k\theta)$$

Note that in the case $k = 0$, we recover a constant velocity on the inner and outer boundaries.

The divergence of an incompressible vector field in polar coordinates is

$$\frac{1}{r}\frac{\partial(r\mathbf{v}_r)}{\partial r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta} = 0$$

or,

$$\frac{\partial(r\mathbf{v}_r)}{\partial r} + \frac{\partial \mathbf{v}_\theta}{\partial \theta} = 0$$

i.e.,

$$\frac{\partial(r\mathbf{v}_r)}{\partial r} = -\frac{\partial \mathbf{v}_\theta}{\partial \theta} = kf(r)\sin(k\theta)$$

so

$$r\mathbf{v}_r(r, \theta) = k\left[\int f(r)dr\right]\sin(k\theta)$$

and finally

$$\mathbf{v}_r(r, \theta) = kg(r)\sin(k\theta)$$

with

$$g(r) = \frac{1}{r}\int f(r)dr \tag{20.26}$$

$$g'(r) = -\frac{1}{r^2}\int f(r)dr + \frac{1}{r}f = -\frac{1}{r}g + \frac{1}{r}f = \frac{1}{r}(f - g) \tag{20.27}$$

The boundary conditions lead to

$$\mathbf{v}_r(r = R_1, \theta) = k\left(\frac{A}{2}R_1 + \frac{B}{R_1}\ln R_1 + \frac{C}{R_1}\right)\sin(k\theta) = 0$$

$$\mathbf{v}_r(r = R_2, \theta) = k\left(\frac{A}{2}R_2 + \frac{B}{R_2}\ln R_2 + \frac{C}{R_2}\right)\sin(k\theta) = 0$$

This has to be valid $\forall\theta$, so

$$\frac{A}{2}R_1^2 + B\ln R_1 = -C \qquad \text{and} \qquad \frac{A}{2}R_2^2 + B\ln R_2 = -C$$

leading to

$$\frac{A}{2} + \frac{B}{R_1^2}\ln R_1 = -\frac{C}{R_1^2} \qquad \text{and} \qquad \frac{A}{2} + \frac{B}{R_2^2}\ln R_2 = -\frac{C}{R_2^2}$$

and finally

$$B = -C\frac{R_2^2 - R_1^2}{R_2^2\ln R_1 - R_1^2\ln R_2}$$

Likewise

$$\frac{A}{2}R_1^2 + B \ln R_1 = -C \qquad \text{and} \qquad \frac{A}{2}R_2^2 + B \ln R_2 = -C$$

yields

$$\frac{A}{2\ln R_1}R_1^2 + B = -\frac{C}{\ln R_1} \qquad \text{and} \qquad \frac{A}{2\ln R_2}R_2^2 + B = -\frac{C}{\ln R_2}$$

or,

$$\frac{A}{2\ln R_1}R_1^2 - \frac{A}{2\ln R_2}R_2^2 = -C(\frac{1}{\ln R_1} - \frac{1}{\ln R_2})$$

$$A(\frac{R_1^2}{2\ln R_1} - \frac{R_2^2}{2\ln R_2}) = -C(\frac{1}{\ln R_1} - \frac{1}{\ln R_2})$$

$$A(\frac{R_1^2 \ln R_2}{2} - \frac{R_2^2 \ln R_1}{2}) = -C(\ln R_2 - \ln R_1)$$

finally

$$A = -C\frac{2(\ln R_2 - \ln R_1)}{R_1^2 \ln R_2 - R_2^2 \ln R_1} = -C\frac{2(\ln R_1 - \ln R_2)}{R_2^2 \ln R_1 - R_1^2 \ln R_2}$$

We set $\vec{g} = -g_r \vec{e}_r$. Stokes equation in Polar coordinates (p284 of Schubert, Turcotte and Olson book):

- r-component:

$$\eta\left[\nabla^2 \mathbf{v}_r - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right] + \frac{\eta}{3}\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial(r\mathbf{v}_r)}{\partial r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right] - \frac{\partial p}{\partial r} - \rho g_r = 0$$

The second term between brackets in the divergence of the velocity field so it is equal to zero in our case. We end up with

$$\eta\left[\nabla^2 \mathbf{v}_r - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta}\right] - \frac{\partial p}{\partial r} - \rho g_r = 0$$

- θ-component:

$$\eta\left[\nabla^2 \mathbf{v}_\theta + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2}\right] + \frac{\eta}{3}\frac{1}{r}\frac{\partial}{\partial \theta}\left[\frac{1}{r}\frac{\partial(r\mathbf{v}_r)}{\partial r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right] - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0$$

The second term between brackets in the divergence of the velocity field so it is equal to zero in our case. We end up with

$$\eta\left[\nabla^2 \mathbf{v}_\theta + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2}\right] - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0$$

In both equations, ∇^2 represents the Laplacian of a scalar quantity:

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}$$

We can then write the two momentum equations for an incompressible Stokes flow in polar coordinates:

$$\eta\left[\frac{\partial^2 \mathbf{v}_r}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_r}{\partial \theta^2} - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta}\right] - \frac{\partial p}{\partial r} - \rho g_r = 0$$

$$\eta\left[\frac{\partial^2 \mathbf{v}_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2}\right] - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0$$

We can further choose $\eta = 1$, so that

$$\frac{\partial^2 \mathbf{v}_r}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_r}{\partial \theta^2} - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta} - \frac{\partial p}{\partial r} - \rho g_r = 0 \qquad (20.28)$$

$$\frac{\partial^2 \mathbf{v}_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2} - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0 \qquad (20.29)$$

Let us define $f(r) = Ar + B/r$. We then have

$$\frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} - \frac{f}{r^2} = A\left(\frac{\partial^2 r}{\partial r^2} + \frac{1}{r}\frac{\partial r}{\partial r} - \frac{1}{r}\right) + B\left(\frac{\partial^2 r^{-1}}{\partial r^2} + \frac{1}{r}\frac{\partial r^{-1}}{\partial r} - \frac{1}{r^3}\right) = 0 \qquad (20.30)$$

Eq. (20.29) simplifies to:

$$\frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0$$

We have

$$\frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} = -k^2\frac{f(r)}{r^2}\cos(k\theta) \qquad \text{and} \qquad \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} = \frac{2k^2}{r^2}g(r)\cos(k\theta)$$

so

$$\frac{1}{r}\frac{\partial p}{\partial \theta} = -\frac{k^2}{r^2}f(r)\cos(k\theta) + \frac{2k^2}{r^2}g(r)\cos(k\theta) = k^2\left(\frac{2g(r) - f(r)}{r^2}\right)\cos(k\theta)$$

and then

$$\frac{\partial p}{\partial \theta} = k^2\left(\frac{2g(r) - f(r)}{r}\right)\cos(k\theta)$$

This can be integrated with respect to θ:

$$p(r,\theta) = k\left(\frac{2g(r) - f(r)}{r}\right)\sin(k\theta) + l(r) = kh(r)\sin(k\theta) + l(r)$$

where $h(r) = \frac{1}{r}(2g(r) - f(r))$. We can turn to Eq.(20.28).

$$
\begin{aligned}
\rho g_r &= \frac{\partial^2 \mathbf{v}_r}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_r}{\partial \theta^2} - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta} - \frac{\partial p}{\partial r} \\
&= +kg''(r)\sin(k\theta) + k\frac{g'(r)}{r}\sin(k\theta) - k^3\frac{g(r)}{r^2}\sin(k\theta) - k\frac{g(r)}{r^2}\sin(k\theta) + k\frac{2f(r)}{r^2}\sin(k\theta) - kh'(r)\sin(k\theta) \\
&= k\sin(k\theta)\left[g'' + \frac{g'}{r} - \frac{k^2 g}{r^2} - \frac{g}{r^2} + \frac{2f}{r^2} - h'\right] - l'(r) \\
&= k\sin(k\theta)\left[g'' + \frac{g'}{r} - \frac{k^2 g}{r^2} - \frac{g}{r^2} + \frac{2f}{r^2} + \frac{1}{r^2}(2g - f) - \frac{1}{r}(2g' - f')\right] - l'(r) \\
&= k\sin(k\theta)\left[g'' + \frac{g'}{r}(1 - 2) - \frac{g}{r^2}(k^2 + 1 - 2) + \frac{f}{r^2}(2 - 1) + \frac{f'}{r}\right] - l'(r) \\
&= k\sin(k\theta)\left[g'' - \frac{g'}{r} - \frac{g}{r^2}(k^2 - 1) + \frac{f}{r^2} + \frac{f'}{r}\right] - l'(r)
\end{aligned}
$$

We can further choose $g_r = 1$. Note that when $k = 0$, we have $\rho = -l'(r)$. We then choose $l'(r) = -\rho_0$ so that the k-dependent term can be seen as a density perturbation:

$$\rho = k\sin(k\theta)\aleph(r) + \rho_0$$

with

$$\aleph(r) = g'' + \frac{g'}{r}\left(1 - \frac{2}{r}\right) - \frac{g}{r^2}\left(k^2 + 1 - \frac{4}{r}\right) + \frac{2f}{r^2}\left(1 - \frac{1}{r}\right) + \frac{f'}{r^2}$$

1046

and

$$f(r) = Ar + \frac{B}{r} \tag{20.32}$$

$$f'(r) = A - \frac{B}{r^2} \tag{20.33}$$

$$g(r) = \frac{A}{2}r + \frac{B}{r}\ln r - \frac{1}{r} \tag{20.34}$$

$$g'(r) = \frac{A}{2} + \frac{B}{r^2}(1 - \ln r) + \frac{1}{r^2} \tag{20.35}$$

$$g''(r) = -\frac{2B}{r^3}(1 - \ln r) - B\frac{1}{r^3} - \frac{2}{r^3} = -\frac{B}{r^3}(3 - 2\ln r) \tag{20.36}$$

Finally, the pressure is then given by

$$p(r, \theta) = k\left(\frac{2g - f}{r^2}\right)\sin(k\theta) + l(r) = kh(r)\sin(k\theta) + \rho_0 g_r r + Constant$$

We enforce $p(r = R_2, \theta) = 0$ so that

$$p(r, \theta) = k\left(\frac{2g - f}{r^2}\right)\sin(k\theta) + l(r) = kh(r)\sin(k\theta) + \rho_0 g_r(r - R_2)$$

Summary of the previous pages :

$$\mathbf{v}_\theta(r, \theta) = f(r)\cos(k\theta) \tag{20.37}$$

$$\mathbf{v}_r(r, \theta) = g(r)k\sin(k\theta) \tag{20.38}$$

$$p(r, \theta) = kh(r)\sin(k\theta) + \rho_0 g_r(r - R_2) \tag{20.39}$$

$$\rho(r, \theta) = k\sin(k\theta)\aleph(r) + \rho_0 \tag{20.40}$$

$$A = \frac{2(\ln R_1 - \ln R_2)}{R_2^2 \ln R_1 - R_1^2 \ln R_2} \tag{20.41}$$

$$B = \frac{R_2^2 - R_1^2}{R_2^2 \ln R_1 - R_1^2 \ln R_2} \tag{20.42}$$

$$f(r) = Ar + \frac{B}{r} \tag{20.43}$$

$$f'(r) = A - \frac{B}{r^2} \tag{20.44}$$

$$g(r) = \frac{A}{2}r + \frac{B}{r}\ln r - \frac{1}{r} \tag{20.45}$$

$$g'(r) = \frac{A}{2} + \frac{B}{r^2}(1 - \ln r) + \frac{1}{r^2} \tag{20.46}$$

$$g''(r) = -\frac{B}{r^3}(3 - 2\ln r) \tag{20.47}$$

$$h(r) = \frac{1}{r^2}(2g - f) \tag{20.48}$$

$$\aleph(r) = -g'' - \frac{g'}{r}\left(1 - \frac{2}{r}\right) + \frac{g}{r^2}\left(k^2 + 1 - \frac{4}{r}\right) - \frac{2f}{r^2}\left(1 - \frac{1}{r}\right) - \frac{f'}{r^2} \tag{20.49}$$

Averagings of fields :

- Average $\mathbf{v}_r$ velocity

$$< \mathbf{v}_r(r) > = \frac{1}{2\pi}\int_0^{2\pi} \mathbf{v}_r(r, \theta)d\theta = \frac{1}{2\pi}\int_0^{2\pi} g(r)k\sin(k\theta)d\theta = \frac{1}{2\pi}g(r)k\int_0^{2\pi}\sin(k\theta)d\theta = 0$$

since $k = 0, 2, 4, ...$

- Average ν_θ velocity

$$< \nu_\theta(r) > = \frac{1}{2\pi} \int_0^{2\pi} \nu_\theta(r, \theta) d\theta = \frac{1}{2\pi} \int_0^{2\pi} f(r) \cos(k\theta) d\theta = \frac{1}{2\pi} f(r) \int_0^{2\pi} \cos(k\theta) d\theta = 0$$

since $k = 0, 2, 4, ...$

- Root mean square verage ν_r velocity

$$
\begin{aligned}
< \nu_r >_{rms}(r) &= \sqrt{\frac{1}{2\pi} \int_0^{2\pi} \nu_r^2 d\theta} \\
&= \sqrt{\frac{1}{2\pi} g(r)^2 k^2 \int_0^{2\pi} \sin^2(k\theta) d\theta} \\
&= \sqrt{\frac{1}{2\pi} g(r)^2 k^2 \int_0^{2\pi} \frac{1}{2}(1 - \cos(2k\theta)) d\theta} \\
&= \sqrt{\frac{1}{2\pi} g(r)^2 k^2 \left(\pi - \frac{1}{2} \int_0^{2\pi} \cos(2k\theta) d\theta \right)} \\
&= \sqrt{\frac{1}{2\pi} g(r)^2 k^2 \left(\pi - \frac{1}{4k} \underbrace{\int_0^{2k\pi} \cos\alpha \, d\alpha}_{=0} \right)} \\
&= \frac{k|g(r)|}{\sqrt{2}}
\end{aligned}
\tag{20.50}
$$

- Root mean square verage ν_θ velocity

$$
\begin{aligned}
< \nu_\theta >_{rms}(r) &= \sqrt{\frac{1}{2\pi} \int_0^{2\pi} \nu_\theta^2 d\theta} \\
&= \sqrt{\frac{1}{2\pi} f(r)^2 \int_0^{2\pi} \cos^2(k\theta) d\theta} \\
&= \sqrt{\frac{1}{2\pi} f(r)^2 \int_0^{2\pi} \frac{1}{2}(1 + \cos(2k\theta)) d\theta} \\
&= \sqrt{\frac{1}{2\pi} f(r)^2 \left(\pi + \frac{1}{2} \int_0^{2\pi} \cos(2k\theta) d\theta \right)} \\
&= \sqrt{\frac{1}{2\pi} f(r)^2 \left(\pi + \frac{1}{4k} \underbrace{\int_0^{4k\pi} \cos\alpha \, d\alpha}_{0} \right)} \\
&= \frac{|f(r)|}{\sqrt{2}}
\end{aligned}
\tag{20.51}
$$

- Root mean square velocity $\mathbf{v}_{rms}$

$$\mathbf{v}_{rms} = \sqrt{\frac{1}{V} \int_V (\mathbf{v}_r^2 + \mathbf{v}_\theta^2)dV} \qquad (20.52)$$

The volume of the domain is given by

$$V = \pi(R_2^2 - R_1^2)$$

and the sum

$$
\begin{aligned}
(\mathbf{v}_r^2 + \mathbf{v}_\theta^2)dV &= [(g(r)k\sin(k\theta))^2 + (f(r)\cos(k\theta))^2]rdrd\theta \\
&= [g(r)^2 rdr][k^2\sin^2(k\theta)d\theta] + [f(r)^2 rdr][\cos^2(k\theta)d\theta]
\end{aligned}
$$

If $k = 0$, we have $\mathbf{v}_\theta = f(r)$ and $\mathbf{v}_r = 0$ so that

$$
\begin{aligned}
\mathbf{v}_{rms} &= \sqrt{\frac{1}{V} \int_0^{2\pi} d\theta \int_{R_1}^{R_2} f(r)^2 rdr} \\
&= \sqrt{\frac{2\pi}{\pi(R_2^2 - R_1^2)} \int_{R_1}^{R_2} f(r)^2 rdr} \\
&= \sqrt{\frac{2}{(R_2^2 - R_1^2)} \int_{R_1}^{R_2} \left(Ar + \frac{B}{r}\right)^2 rdr} \\
&= \sqrt{\frac{2}{(R_2^2 - R_1^2)} \int_{R_1}^{R_2} \left(A^2 r^3 + 2ABr + \frac{B^2}{r}\right) dr} \\
&= \sqrt{\frac{2}{(R_2^2 - R_1^2)} \left[A^2\frac{r^4}{4} + ABr^2 + B^2\ln(r)\right]_{R_1}^{R_2}} \\
&= \sqrt{\frac{2}{(R_2^2 - R_1^2)}[\frac{A^2}{4}(R_2^4 - R_1^4) + AB(R_2^2 - R_1^2) + B^2(\ln R_2 - \ln R_1)]} \quad (20.53)
\end{aligned}
$$

If $k \neq 0$ we have

$$\int_0^{2\pi} k^2\sin^2(k\theta)d\theta = \frac{1}{2}k^2 \int_0^{2\pi} (1 - \cos(k\theta))d\theta = \pi k^2$$

$$\int_0^{2\pi} \cos^2(k\theta)d\theta = \frac{1}{2}\int_0^{2\pi} (1 + \cos(k\theta))d\theta = \pi$$

so that

$$\mathcal{V} = \int_0^{2\pi}\int_{R_1}^{R_2} (\mathbf{v}_r^2 + \mathbf{v}_\theta^2)rdrd\theta = \pi k^2 \int_{R_1}^{R_2} g(r)^2 rdr + \pi \int_{R_1}^{R_2} f(r)^2 rdr$$

$$\int_{R_1}^{R_2} f(r)^2 r dr = \int_{R_1}^{R_2} \left(Ar + \frac{B}{r} \right)^2 r dr$$

$$= \int_{R_1}^{R_2} \left(A^2 r^3 + 2ABr + \frac{B^2}{r} \right) dr$$

$$= \left[A^2 \frac{r^4}{4} + ABr^2 + B^2 \ln(r) \right]_{R_1}^{R_2}$$

$$= \frac{A^2}{4} (R_2^4 - R_1^4) + AB(R_2^2 - R_1^2) + B^2 (\ln R_2 - \ln R_1)$$

$$\int_{R_1}^{R_2} g(r)^2 r dr = \int_{R_1}^{R_2} \left(\frac{A}{2} r + \frac{B}{r} \ln r + \frac{C}{r} \right)^2 r dr$$

$$= \int_{R_1}^{R_2} \left(\frac{A^2}{4} r^2 + AB \ln r + AC + \frac{2BC}{r^2} \ln r + \frac{B^2}{r^2} (\ln r)^2 + \frac{C^2}{r^2} \right) r dr$$

$$= \int_{R_1}^{R_2} \left(\frac{A^2}{4} r^3 + ABr \ln r + ACr + \frac{2BC}{r} \ln r + \frac{B^2}{r} (\ln r)^2 + \frac{C^2}{r} \right) dr$$

$$= \int_{R_1}^{R_2} \left(\frac{A^2}{4} r^3 + ACr + \frac{C^2}{r} \right) dr + E + F + G$$

$$= \left[\frac{A^2}{16} r^4 + \frac{1}{2} ACr^2 + C^2 \ln r \right]_{R_1}^{R_2} + E + F + G$$

$$= \frac{A^2}{16} (R_2^4 - R_1^4) + \frac{AC}{2} (R_2^2 - R_1^2) + C^2 (\ln R_2 - \ln R_1) + E + F + G$$

$$(20.54)$$

$$
\begin{aligned}
E &= 2BC \int_{R_1}^{R_2} \frac{1}{r} \ln r \, dr \\
&= BC \int_{\ln R_1}^{\ln R_2} 2X dX \qquad X = \ln r, \;\; dX = dr/r \\
&= BC[X^2]_{\ln R_1}^{\ln R_2} \\
&= BC[(\ln R_2)^2 - (\ln R_1)^2]
\end{aligned}
$$

$$
\begin{aligned}
F &= B^2 \int_{R_1}^{R_2} \frac{1}{r}(\ln r)^2 dr \\
&= B^2 \int_{\ln R_1}^{\ln R_2} X^2 dX \qquad X = \ln r, \;\; dX = dr/r \\
&= \frac{B^2}{3}[X^3]_{\ln R_1}^{\ln R_2} \\
&= \frac{B^2}{3}[(\ln R_2)^3 - (\ln R_1)^3]
\end{aligned}
$$

$$
\begin{aligned}
G &= AB \int_{R_1}^{R_2} r \ln r \, dr \\
&= AB[\frac{1}{2}r^2 \ln r]_{R_1}^{R_2} - AB \int_{R_1}^{R_2} \frac{1}{2}r^2 \frac{1}{r} dr \\
&= \frac{AB}{2}[R_2^2 \ln R_2 - R_1^2 \ln R_1] - \frac{AB}{2} \int_{R_1}^{R_2} r \, dr \\
&= \frac{AB}{2}[R_2^2 \ln R_2 - R_1^2 \ln R_1] - \frac{AB}{4}[r^2]_{R_1}^{R_2} \\
&= \frac{AB}{2}[R_2^2 \ln R_2 - R_1^2 \ln R_1] - \frac{AB}{4}(R_2^2 - R_1^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{V}/\pi &= \frac{A^2}{4}(R_2^4 - R_1^4) + AB(R_2^2 - R_1^2) + B^2(\ln R_2 - \ln R_1) \\
&+ \frac{A^2 k^2}{16}(R_2^4 - R_1^4) + \frac{ACk^2}{2}(R_2^2 - R_1^2) + C^2 k^2(\ln R_2 - \ln R_1) \\
&+ BCk^2[(\ln R_2)^2 - (\ln R_1)^2] \\
&+ \frac{B^2 k^2}{3}[(\ln R_2)^3 - (\ln R_1)^3] \\
&+ \frac{ABk^2}{2}[R_2^2 \ln R_2 - R_1^2 \ln R_1] - \frac{ABk^2}{4}(R_2^2 - R_1^2) \\
&= \frac{A^2}{16}(4 + k^2)(R_2^4 - R_1^4) + (\frac{AB}{4}(4 - k^2) + \frac{ACk^2}{2})(R_2^2 - R_1^2) \\
&+ (B^2 + C^2 k^2)(\ln R_2 - \ln R_1) \\
&+ BCk^2[(\ln R_2)^2 - (\ln R_1)^2] \\
&+ \frac{B^2 k^2}{3}[(\ln R_2)^3 - (\ln R_1)^3] \\
&+ \frac{ABk^2}{2}[R_2^2 \ln R_2 - R_1^2 \ln R_1] \qquad\qquad (20.55)
\end{aligned}
$$

$$\mathbf{v}_{rms} = \sqrt{\frac{1}{V}\mathcal{V}} = \sqrt{\frac{1}{2(R_2^2 - R_1^2)}\frac{\mathcal{V}}{\pi}}$$

Computing the strain rate and stress tensors :

Since we know the viscosity, the velocity field and the pressure field, we can also compute the full stress tensor $\boldsymbol{\sigma} = -p\mathbf{1} + 2\eta\dot{\boldsymbol{\varepsilon}}$ In this benchmark we have set $\eta = 1$ so: $\boldsymbol{\sigma} = -p\mathbf{1} + 2\dot{\boldsymbol{\varepsilon}}$ We start with the velocity gradient:

$$
\begin{aligned}
\vec{\nabla}\vec{v} &= \begin{pmatrix} \frac{\partial v_r}{\partial r} & \frac{1}{r}\frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r} \\[2mm] \frac{\partial v_\theta}{\partial r} & \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r} \end{pmatrix} \\[2mm]
&= \begin{pmatrix} g'k\sin(k\theta) & \frac{1}{r}gk^2\cos(k\theta) - \frac{1}{r}f\cos(k\theta) \\[2mm] f'\cos(k\theta) & -\frac{1}{r}fk\sin(k\theta) + \frac{1}{r}gk\sin(k\theta) \end{pmatrix} \\[2mm]
&= \begin{pmatrix} g'k\sin(k\theta) & \frac{1}{r}(gk^2 - f)\cos(k\theta) \\[2mm] f'\cos(k\theta) & \frac{1}{r}(g - f)k\sin(k\theta) \end{pmatrix}
\end{aligned}
\tag{20.56}
$$

The strain rate is then given by:

$$
\dot{\boldsymbol{\varepsilon}} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{v} + \boldsymbol{\nabla}\boldsymbol{v}^T) = \begin{pmatrix} g'k\sin(k\theta) & \frac{1}{2r}(rf' + gk^2 - f)\cos(k\theta) \\[2mm] \frac{1}{2r}(rf' + gk^2 - f)\cos(k\theta) & \frac{1}{r}(g - f)k\sin(k\theta) \end{pmatrix}
\tag{20.57}
$$

Let us verify once again that the flow is incompressible:

$$
\vec{\nabla}\cdot\vec{v} = g'(r)k\sin(k\theta) + \frac{1}{r}(g(r) - f(r))k\sin(k\theta) = \frac{1}{r}(rg'(r) + g(r) - f(r))k\sin(k\theta) = 0
$$

since $g'(r) = \frac{1}{r}(f(r) - g(r))$.

I can now write the full stress tensor:

$$
\boldsymbol{\sigma} = -p\mathbf{1} + 2\dot{\boldsymbol{\varepsilon}} = \begin{pmatrix} -p + 2g'k\sin(k\theta) & \frac{1}{r}(rf' + gk^2 - f)\cos(k\theta) \\[2mm] \frac{1}{r}(rf' + gk^2 - f)\cos(k\theta) & -p + \frac{2}{r}(g - f)k\sin(k\theta) \end{pmatrix}
\tag{20.58}
$$

On the boundaries, i.e. $r = R_1$ or $r = R_2$, the function g is exactly zero, so that the stress $\boldsymbol{\sigma}_b$ on the boundaries is given by

$$
\boldsymbol{\sigma}_b = \begin{pmatrix} -p + 2g'k\sin(k\theta) & \frac{1}{r}(rf' - f)\cos(k\theta) \\[2mm] \frac{1}{r}(rf' - f)\cos(k\theta) & -p - \frac{2}{r}fk\sin(k\theta) \end{pmatrix}
\tag{20.59}
$$

Furthermore I can use the identity $g' = (f - g)/r$ which simplifies to $g' = f/r$ on the boundaries:

$$
\boldsymbol{\sigma}_b = \begin{pmatrix} -p + 2\frac{f}{r}k\sin(k\theta) & \frac{1}{r}(rf' - f)\cos(k\theta) \\[2mm] \frac{1}{r}(rf' - f)\cos(k\theta) & -p - \frac{2}{r}fk\sin(k\theta) \end{pmatrix}
\tag{20.60}
$$

Also, $h(r) = \frac{1}{r}(2g - f)$ simplifies to $h(r) = -f/r$ so

$$
p = kh(r)\sin(k\theta) + \rho_0 g_r(r - R_2) = -k\frac{f}{r}\sin(k\theta) + \rho_0 g_r(r - R_2)
$$

Finally

$$
\boldsymbol{\sigma_b} = \begin{pmatrix} k\frac{f}{r}\sin(k\theta) - \rho_0 g_r(r - R_2) + 2\frac{f}{r}k\sin(k\theta) & \frac{1}{r}(rf' - f)\cos(k\theta) \\ \frac{1}{r}(rf' - f)\cos(k\theta) & k\frac{f}{r}\sin(k\theta) - \rho_0 g_r(r - R_2) - \frac{2}{r}fk\sin(k\theta) \end{pmatrix}
$$

$$
= \begin{pmatrix} k\frac{3f}{r}\sin(k\theta) - \rho_0 g_r(r - R_2) & \frac{1}{r}(rf' - f)\cos(k\theta) \\ \frac{1}{r}(rf' - f)\cos(k\theta) & -k\frac{f}{r}\sin(k\theta) - \rho_0 g_r(r - R_2) \end{pmatrix}
$$

The traction along the normal is given by

$$
\boldsymbol{\sigma}_n = \boldsymbol{\sigma}_b \cdot \vec{n} = -\boldsymbol{\sigma}_b \cdot \vec{e_r} = - \begin{pmatrix} k\frac{3f}{r}\sin(k\theta) - \rho_0 g_r(r - R_2) \\ \frac{1}{r}(rf' - f)\cos(k\theta) \end{pmatrix}
$$

Strain rate tensor in Cartesian coordinates The analytical expressions for the strain rate components in polar coordinates are:

$$
\dot{\varepsilon}_{rr} = g'(r)k\sin k\theta \tag{20.61}
$$

$$
\dot{\varepsilon}_{r\theta} = \dot{\varepsilon}_{\theta r} = \frac{1}{2}\left(\frac{1}{r}g(r)k^2\cos k\theta + f'(r)\cos k\theta - \frac{f(r)}{r}\cos k\theta\right) \tag{20.62}
$$

$$
= \frac{1}{2}\left(\frac{1}{r}g(r)k^2 + f'(r) - \frac{f(r)}{r}\right)\cos k\theta \tag{20.63}
$$

$$
\dot{\varepsilon}_{\theta\theta} = -\frac{1}{r}kf(r)\sin k\theta + \frac{g(r)}{r}k\sin k\theta \tag{20.64}
$$

$$
= \frac{g(r) - f(r)}{r}k\sin k\theta \tag{20.65}
$$

Their counterparts in Cartesian coordinates are obtained with Eqs. (4.27).

Could we find a steady state temperature field that goes along? We can start from the pure advection equation:

$$
\rho C_p\left(\frac{\partial T}{\partial t} + \vec{v}\cdot\vec{\nabla}T\right) = 0
$$

At steady state we are left with

$$
\vec{v}\cdot\vec{\nabla}T = 0
$$

I postulate then

$$
T(r,\theta) = l(r)(\alpha\cos k\theta + \beta\sin k\theta)
$$

Inserting this into the equation above I arrive at the conclusion that $\beta \neq 0$ is a dead end. It then simply follows that

$$
T(r,\theta) = l(r)\cos k\theta
$$

which yields $T(r,\theta) = rg(r)\cos(k\theta)$. and

$$
\vec{\nabla}T = (f(r)\cos k\theta; -g(r)k\sin k\theta)
$$

I can then plot the temperature gradient (in black) next to the velocity field in red (right figure) and show that these are indeed always perpendicular:
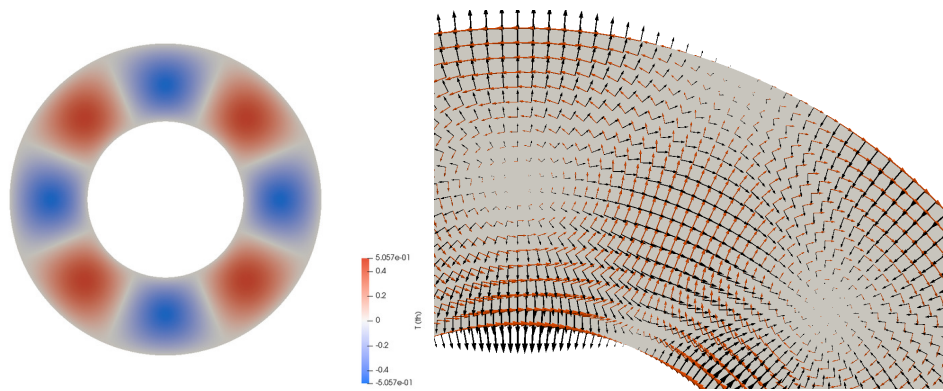
However, what is deeply puzzling is the fact that the temperature is exactly zero on both boundaries (beacuse $g(r)$ is by construction) and yet the temperature field is not zero in the middle, in the absence of heat source ... ?

20.19   The annulus convection benchmark #2 - no slip bc

We seek an exact solution to the incompressible Stokes equations for an isoviscous, isothermal fluid in an annulus. Given the geometry of the problem, we work in polar coordinates. We denote the orthonormal basis vectors by $\vec{e}_r$ and $\vec{e}_\theta$, the inner radius of the annulus by R_1 and the outer radius by R_2. Further, we assume that the viscosity η is constant, which we set to $\eta = 1$, we set the gravity vector to $\vec{g} = -g_r\,\vec{e}_r$ with $g_r = 1$.

Given these assumptions, the incompressible Stokes equations in the annulus are (see Schubert, Turcotte & Olson [1216]):

$$A_r = \frac{\partial^2 \boldsymbol{\nu}_r}{\partial r^2} + \frac{1}{r}\frac{\partial \boldsymbol{\nu}_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \boldsymbol{\nu}_r}{\partial \theta^2} - \frac{\boldsymbol{\nu}_r}{r^2} - \frac{2}{r^2}\frac{\partial \boldsymbol{\nu}_\theta}{\partial \theta} - \frac{\partial p}{\partial r} \;\; = \;\; \rho g_r \qquad (20.66)$$

$$A_\theta = \frac{\partial^2 \boldsymbol{\nu}_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial \boldsymbol{\nu}_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \boldsymbol{\nu}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \boldsymbol{\nu}_r}{\partial \theta} - \frac{\boldsymbol{\nu}_\theta}{r^2} - \frac{1}{r}\frac{\partial p}{\partial \theta} \;\; = \;\; 0 \qquad (20.67)$$

$$\frac{1}{r}\frac{\partial(r\boldsymbol{\nu}_r)}{\partial r} + \frac{1}{r}\frac{\partial \boldsymbol{\nu}_\theta}{\partial \theta} \;\; = \;\; 0 \qquad (20.68)$$

Equations (20.66) and (20.67) are the momentum equations in polar coordinates while Equation (20.68) is the mass conservation equation (also called continuity equation). The components of the velocity are obtained from the stream function Ψ as follows:

$$\boldsymbol{\nu}_r = \frac{1}{r}\frac{\partial \Psi}{\partial \theta} \qquad \boldsymbol{\nu}_\theta = -\frac{\partial \Psi}{\partial r}$$

where $\boldsymbol{\nu}_r$ is the radial component and $\boldsymbol{\nu}_\theta$ is the tangential component of the velocity vector.

The stream function is defined for incompressible (divergence-free) flows in 2D (as well as in 3D with axisymmetry). The stream function can be used to plot streamlines, which represent the trajectories of particles in a steady flow. From calculus it is known that the gradient vector $\nabla\Psi$ is normal to the curve $\Psi = C$. It can be shown that everywhere $\vec{\nu} \cdot \nabla\Psi = 0$ using the formula for $\vec{u}$ in terms of Ψ which proves that level curves of Ψ are streamlines:

$$\vec{\boldsymbol{\nu}} \cdot \nabla\Psi = \boldsymbol{\nu}_r\frac{\partial \Psi}{\partial r} + \boldsymbol{\nu}_\theta\frac{1}{r}\frac{\partial \Psi}{\partial \theta} = \frac{1}{r}\frac{\partial \Psi}{\partial \theta}\frac{\partial \Psi}{\partial r} - \frac{\partial \Psi}{\partial r}\frac{1}{r}\frac{\partial \Psi}{\partial \theta} = 0$$

In polar coordinates the curl of a vector $\vec{A}$ is [2]:

$$\vec{\nabla} \times \vec{A} = \frac{1}{r}\left(\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta}\right)\vec{e}_z$$

Taking the curl of vector $\vec{A}$ (see Eqs. (20.66) and (20.67)) yields:

$$\frac{1}{r}\left(\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta}\right) = \frac{1}{r}\left(-\frac{\partial(\rho g_r)}{\partial \theta}\right)$$

Multiplying on each side by r

$$\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta} = -\frac{\partial \rho g_r}{\partial \theta}$$

If we now replace A_r and A_θ by their expressions as a function of velocity and pressure, we will see that the pressure terms cancel out (which is one of the advantages of working with stream line formulations).

Let us assume the following separation of variables $\boxed{\Psi(r,\theta) = \phi(r)\xi(\theta)}$. Then

$$\boldsymbol{\nu}_r = \frac{1}{r}\frac{\partial \Psi}{\partial \theta} = \frac{\phi\xi'}{r} \qquad \boldsymbol{\nu}_\theta = -\frac{\partial \Psi}{\partial r} = -\phi'\xi$$

---

[2] https://en.wikipedia.org/wiki/Del_in_cylindrical_and_spherical_coordinates

Let us first express A_r and A_θ as functions of ϕ and ξ:

$$
\begin{aligned}
A_r &= \frac{\partial^2 \mathbf{v}_r}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_r}{\partial \theta^2} - \frac{\mathbf{v}_r}{r^2} - \frac{2}{r^2}\frac{\partial \mathbf{v}_\theta}{\partial \theta} - \frac{\partial p}{\partial r} \\
&= \frac{\partial^2}{\partial r^2}\left(\frac{\phi\xi'}{r}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{\phi\xi'}{r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\left(\frac{\phi\xi'}{r}\right) - \frac{1}{r^2}\left(\frac{\phi\xi'}{r}\right) - \frac{2}{r^2}\frac{\partial}{\partial \theta}(-\phi'\xi) - \frac{\partial p}{\partial r} \\
&= \left(\frac{\phi''}{r} - 2\frac{\phi'}{r^2} + 2\frac{\phi}{r^3}\right)\xi' + \left(\frac{\phi'}{r^2} - \frac{\phi}{r^3}\right)\xi' + \frac{\phi}{r^3}\xi''' - \frac{\phi\xi'}{r^3} + \frac{2}{r^2}\phi'\xi' - \frac{\partial p}{\partial r} \\
&= \frac{\phi''\xi'}{r} + \frac{\phi'\xi'}{r^2} + \frac{\phi\xi'''}{r^3} - \frac{\partial p}{\partial r} \\
\frac{\partial A_r}{\partial \theta} &= \frac{\phi''\xi''}{r} + \frac{\phi'\xi''}{r^2} + \frac{\phi\xi''''}{r^3} - \frac{\partial^2 p}{\partial r \partial \theta} \qquad\qquad\qquad\qquad\qquad\qquad (20.69)
\end{aligned}
$$

$$
\begin{aligned}
A_\theta &= \frac{\partial^2 \mathbf{v}_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2} - \frac{1}{r}\frac{\partial p}{\partial \theta} \\
&= \frac{\partial^2}{\partial r^2}(-\phi'\xi) + \frac{1}{r}\frac{\partial}{\partial r}(-\phi'\xi) + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}(-\phi'\xi) + \frac{2}{r^2}\frac{\partial}{\partial \theta}\left(\frac{\phi\xi'}{r}\right) - \frac{1}{r^2}(-\phi'\xi) - \frac{1}{r}\frac{\partial p}{\partial \theta} \\
&= -\phi'''\xi - \frac{\phi''\xi}{r} + \frac{\phi'(\xi - \xi'')}{r^2} + \frac{2\phi\xi''}{r^3} - \frac{1}{r}\frac{\partial p}{\partial \theta} \\
r A_\theta &= -\phi'''\xi r - \phi''\xi + \frac{\phi'(\xi - \xi'')}{r} + \frac{2\phi\xi''}{r^2} - \frac{\partial p}{\partial \theta} \\
\frac{\partial(r A_\theta)}{\partial r} &= -(\phi'''\xi + \phi''''\xi r) - \phi'''\xi + (\xi - \xi'')\left(\frac{\phi''}{r} - \frac{\phi'}{r^2}\right) + \left(\frac{2\phi'\xi''}{r^2} - 2\frac{2\phi\xi''}{r^3}\right) - \frac{\partial^2 p}{\partial \theta \partial r} \\
&= -2\phi'''\xi - \phi''''\xi r + \frac{1}{r}\phi''(\xi - \xi'') + \frac{1}{r^2}(-\phi'(\xi - \xi'') + 2\phi'\xi'') + \frac{1}{r^3}(-4\phi\xi'') - \frac{\partial^2 p}{\partial \theta \partial r} \\
&= -2\phi'''\xi - \phi''''\xi r + \frac{\phi''}{r}(\xi - \xi'') + \frac{\phi'}{r^2}(-\xi + 3\xi'') + \frac{\phi}{r^3}(-4\xi'') - \frac{\partial^2 p}{\partial \theta \partial r} \qquad (20.70)
\end{aligned}
$$

No slip boundary conditions

No-slip boundary conditions inside and outside impose that all components of the velocity must be zero on both boundaries, i.e.

$$\vec{\mathbf{v}}(r = R_1, \theta) = \vec{\mathbf{v}}(r = R_2, \theta) = \vec{0}$$

Due to the separation of variables, and choosing $\boxed{\xi(\theta) = \cos(k\theta)}$ we have

$$
\begin{aligned}
\mathbf{v}_r(r, \theta) &= \frac{1}{r}\frac{\partial \Psi}{\partial \theta} = \frac{\phi\xi'}{r} = -\frac{1}{r}\phi(r)k\sin(k\theta) & (20.71) \\
\mathbf{v}_\theta(r, \theta) &= -\frac{\partial \Psi}{\partial r} = -\phi'(r)\xi(\theta) = -\phi'(r)\cos(k\theta) & (20.72)
\end{aligned}
$$

The velocity divergence is given by

$$\vec{\nabla}\cdot\vec{\mathbf{v}} = \frac{1}{r}\frac{\partial(r\mathbf{v}_r)}{\partial r} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial \theta} = \frac{1}{r}\left(-\phi'(r)k\sin(k\theta) + \phi'(r)k\sin(k\theta)\right) = 0$$

so the flow is indeed incompressible.

Since ξ is a function of θ it is obvious that the only way to insure no-slip boundary conditions for any θ value is to have all the following four conditions satisfied

$$
\begin{aligned}
\phi(R_1) = \phi'(R_1) &= 0 & (20.73) \\
\phi(R_2) = \phi'(R_2) &= 0 & (20.74)
\end{aligned}
$$

We could then choose

$$\phi(r) = (r - R_1)^2(r - R_2)^2 \mathcal{F}(r) \tag{20.75}$$
$$\phi'(r) = 2(r - R_1)(r - R_2)^2 \mathcal{F}(r) + 2(r - R_1)^2(r - R_2)\mathcal{F}(r) + (r - R_1)^2(r - R_2)^2 \mathcal{F}'(r) \tag{20.76}$$

which are indeed identically zero on both boundaries. Here $\mathcal{F}(r)$ is any (smooth enough) function of r. A generic form for Ψ could then be

$$\boxed{\Psi(r, \theta) = (r - R_1)^2(r - R_2)^2 \mathcal{F}(r) \cos(k\theta)}$$

In what follows I take $\mathcal{F}(r) = 1$ for simplicity. Then

$$
\begin{aligned}
\phi(r) &= (r - R_1)^2(r - R_2)^2 \\
&= (r^2 - 2rR_1 + R_1^2)(r^2 - 2rR_2 + R_2^2) \\
&= \underbrace{1}_{a} r^4 + \underbrace{(-2R_1 - 2R_2)}_{b} r^3 + \underbrace{(R_1^2 + R_2^2 + 4R_1 R_2)}_{c} r^2 + \underbrace{(-2R_1 R_2^2 - 2R_1^2 R_2)}_{d} r + \underbrace{R_1^2 R_2^2}_{e} \\
&= ar^4 + br^3 + cr^2 + dr + e \tag{20.77}
\end{aligned}
$$

and then

$$
\begin{aligned}
\phi'(r) &= 2(r - R_1)(r - R_2)^2 + 2(r - R_1)^2(r - R_2) \tag{20.78} \\
&= 2(r - R_1)(r - R_2)(r - R_2 + r - R_1) \tag{20.79} \\
&= 4(r - R_1)(r - R_2)\left(r - \frac{R_1 + R_2}{2}\right) \tag{20.80} \\
\phi''(r) &= 8\left(r - \frac{R_1 + R_2}{2}\right)^2 + 4(r - R_1)(r - R_2) \tag{20.81} \\
\phi'''(r) &= 24\left(r - \frac{R_1 + R_2}{2}\right) \tag{20.82} \\
\phi''''(r) &= 24 \tag{20.83} \\
v_r(r, \theta) &= -\frac{1}{r}(r - R_1)^2(r - R_2)^2 \, k \sin(k\theta) \tag{20.84} \\
v_\theta(r, \theta) &= -4(r - R_1)(r - R_2)\left(r - \frac{R_1 + R_2}{2}\right)\cos(k\theta) \tag{20.85}
\end{aligned}
$$

In the end the functions ϕ and ξ are of the form:

$$
\begin{aligned}
\xi(\theta) &= \cos(k\theta) \tag{20.86} \\
\phi(r) &= ar^4 + br^3 + cr^2 + dr + e \tag{20.87}
\end{aligned}
$$

with

$$
\begin{aligned}
\xi'(\theta) &= -k\sin(k\theta) \\
\xi''(\theta) &= -k^2 \cos(k\theta) = -k^2 \xi(\theta) \\
\xi'''(\theta) &= k^3 \sin(k\theta) \\
\xi''''(\theta) &= k^4 \cos(k\theta) = k^4 \xi(\theta) \\
\phi'(r) &= 4ar^3 + 3br^2 + 2cr + d \\
\phi''(r) &= 12ar^2 + 6br + 2c \\
\phi'''(r) &= 24ar + 6b \\
\phi''''(r) &= 24a
\end{aligned}
$$

Finding the pressure and density fields

We start from the relationship

$$A_\theta = \frac{\partial^2 \mathbf{v}_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{v}_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathbf{v}_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial \mathbf{v}_r}{\partial \theta} - \frac{\mathbf{v}_\theta}{r^2} - \frac{1}{r}\frac{\partial p}{\partial \theta} = -\phi'''\xi - \frac{\phi''\xi}{r} + \frac{\phi'(\xi - \xi'')}{r^2} + \frac{2\phi\xi''}{r^3} - \frac{1}{r}\frac{\partial p}{\partial \theta} = 0$$

Then, after multiplying all by r^3 we have

$$r^2\frac{\partial p}{\partial \theta}$$
$$= -r^3\phi'''\xi - r^2\phi''\xi + r\phi'(\xi - \xi'') + 2\phi\xi''$$
$$= \cos(k\theta)\left[-r^3(24ar + 6b) - r^2(12ar^2 + 6br + 2c) + r(4ar^3 + 3br^2 + 2cr + d)(1 + k^2) + 2(ar^4 + br^3 + cr^2\right.$$
$$= \cos(k\theta)\left[-24ar^4 - 6br^3 - 12ar^4 - 6br^3 - 2cr^2 + (4ar^4 + 3br^3 + 2cr^2 + dr)(1 + k^2) - 2k^2(ar^4 + br^3 + cr^2\right.$$
$$= \cos(k\theta)\left[(-24 - 12 + 4 + 4k^2 - 2k^2)ar^4 + (-6 - 6 + 3 + 3k^2 - 2k^2)br^3 + (-2 + 2 + 2k^2 - 2k^2)cr^2 + (1\right.$$
$$= \cos(k\theta)\left[2(k^2 - 16)ar^4 + (k^2 - 9)br^3 + (1 - k^2)dr - 2k^2e\right]$$

or,

$$\frac{\partial p}{\partial \theta} = \cos(k\theta)\left[2(k^2 - 16)ar^2 + (k^2 - 9)br + (1 - k^2)\frac{d}{r} - 2k^2\frac{e}{r^2}\right]$$

i.e. after integration with respect to θ:

$$p(r, \theta) = \frac{1}{k}\sin(k\theta)\left[2(k^2 - 16)ar^2 + (k^2 - 9)br + (1 - k^2)\frac{d}{r} - 2k^2\frac{e}{r^2}\right] + f(r)$$

For simplicity we set $f(r) = 0$ and then

$$\frac{\partial p}{\partial r} = \frac{1}{k}\sin(k\theta)\left[4(k^2 - 16)ar + (k^2 - 9)b - (1 - k^2)\frac{d}{r^2} + 4k^2\frac{e}{r^3}\right]$$

and since we will need it later:

$$r^3\frac{\partial p}{\partial r} = \frac{1}{k}\sin(k\theta)\left[4(k^2 - 16)ar^4 + (k^2 - 9)br^3 - (1 - k^2)dr + 4k^2e\right]$$

We now turn to

$$A_r = \frac{\phi''\xi'}{r} + \frac{\phi'\xi'}{r^2} + \frac{\phi\xi'''}{r^3} - \frac{\partial p}{\partial r} = \rho g_r$$

or, after multiplying both sides by r^3:

$$r^2\phi''\xi' + r\phi'\xi' + \phi\xi''' - r^3\frac{\partial}{\partial}$$

$$r^2(12ar^2 + 6br + 2c)(-k\sin(k\theta)) + r(4ar^3 + 3br^2 + 2cr + d)(-k\sin(k\theta)) + (ar^4 + br^3 + cr^2 + dr + e)k^3\sin(k\theta)$$

$$-\frac{1}{k}\sin(k\theta)\left[4(k^2 - 16)ar^4 + (k^2 - 9)br^3 - (1 - k^2)dr + 4k^2e\right.$$

$$\sin(k\theta)\left[-k^2(12ar^4 + 6br^3 + 2cr^2) - k^2(4ar^4 + 3br^3 + 2cr + dr) + k^4(ar^4 + br^3 + cr^2 + dr + e)\right.$$

$$-\sin(k\theta)\left[4(k^2 - 16)ar^4 + (k^2 - 9)br^3 - (1 - k^2)dr + 4k^2e\right.$$

$$\sin(k\theta)\left[(-12k^2 - 4k^2 + k^4 - 4(k^2 - 16))ar^4 + (-6k^2 - 3k^2 + k^4 - (k^2 - 9))br\right.$$
$$(-2k^2 - 2k^2 + k^4)cr^2 + (-k^2 + k^4 + (1 - k^2))dr + (k^4 - 4k^2e)$$

$$\sin(k\theta)\left[(k^4 - 20k^2 + 64)ar^4 + (k^4 - 10k^2 + 9)br^3 + (k^4 - 4k^2)cr^2 + (k^4 - 2k^2 + 1)dr + k^2(k^2 - 4)e\right.$$

So, assuming $g_r = 1$,

$$\rho(r, \theta) = \frac{\sin(k\theta)}{k}\frac{Aar^4 + Bbr^3 + Ccr^2 + Ddr + Ee}{r^3}$$

with

$$
\begin{aligned}
A &= (k^2 - 4)(k^2 - 16) \\
B &= k^4 - 10k^2 + 9 \\
C &= k^2(k^2 - 4) \\
D &= (k^2 - 1)(k^2 - 1) \\
E &= k^2(k^2 - 4)
\end{aligned}
\tag{20.92}
$$

In the end:

$$
\mathbf{v}_r(r, \theta) = -\frac{1}{r}(ar^4 + br^3 + cr^2 + dr + e)k\sin(k\theta) \tag{20.93}
$$

$$
\mathbf{v}_\theta(r, \theta) = -(4ar^3 + 3br^2 + 2cr + d)\cos(k\theta) \tag{20.94}
$$

$$
p(r, \theta) = \frac{1}{k}\sin(k\theta)\left[2(k^2 - 16)ar^2 + (k^2 - 9)br + (1 - k^2)\frac{d}{r} - 2k^2\frac{e}{r^2}\right] \tag{20.95}
$$

$$
\rho(r, \theta) = \frac{\sin(k\theta)}{k}\frac{Aar^4 + Bbr^3 + Ccr^2 + Ddr + Ee}{r^3} \tag{20.96}
$$

Finding the density field - alternate & easier take

We start this time from
$$
\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta} = -\frac{\partial \rho g_r}{\partial \theta}
$$

or, with $g_r = 1$,

$$
\begin{aligned}
-2\phi'''\xi - \phi''''\xi r + \frac{\phi''}{r}(\xi - \xi'') + \frac{\phi'}{r^2}(-\xi + 3\xi'') + \frac{\phi}{r^3}(-4\xi'') - \frac{\partial^2 p}{\partial\theta\partial r} - \frac{\phi''\xi''}{r} - \frac{\phi'\xi''}{r^2} - \frac{\phi\xi''''}{r^3} + \frac{\partial^2 p}{\partial r\partial\theta} &= -\frac{\partial\rho}{\partial\theta} \\
-2\phi'''\xi - \phi''''\xi r + \frac{\phi''}{r}(\xi - \xi'') + \frac{\phi'}{r^2}(-\xi + 3\xi'') + \frac{\phi}{r^3}(-4\xi'') - \frac{\phi''\xi''}{r} - \frac{\phi'\xi''}{r^2} - \frac{\phi\xi''''}{r^3} &= -\frac{\partial\rho}{\partial\theta}
\end{aligned}
$$

Then we note that $\xi'' = -k^2\xi$ and $\xi'''' = k^4\xi$ so that

$$
\begin{aligned}
-2\phi'''\xi - \phi''''\xi r + \frac{\phi''}{r}(\xi + k^2\xi) + \frac{\phi'}{r^2}(-\xi - 3k^2\xi) + \frac{\phi}{r^3}(4k^2\xi) - \frac{-k^2\phi''\xi}{r} - \frac{-k^2\phi'\xi}{r^2} - \frac{k^4\phi\xi}{r^3} &= -\frac{\partial\rho}{\partial\theta} \\
\xi\left[-2\phi''' - \phi''''r + \frac{\phi''}{r}(1 + k^2) + \frac{\phi'}{r^2}(-1 - 3k^2) + \frac{\phi}{r^3}(4k^2) - \frac{-k^2\phi''}{r} - \frac{-k^2\phi'}{r^2} - \frac{k^4\phi}{r^3}\right] &= -\frac{\partial\rho}{\partial\theta} \\
\xi(\theta)\left[-2\phi''' - \phi''''r + \frac{\phi''}{r}(1 + 2k^2) + \frac{\phi'}{r^2}(-1 - 2k^2) + \frac{\phi}{r^3}(4k^2 - k^4)\right] &= -\frac{\partial\rho}{\partial\theta}
\end{aligned}
$$

i.e.

$$
\rho(r, \theta) = -\frac{1}{k}\sin(k\theta)\frac{\mathcal{G}(r)}{r^3}
$$

with

$$
\begin{aligned}
\mathcal{G}(r) &= -\phi''''r^4 - 2\phi'''r^3 + \phi''r^2(1 + 2k^2) + \phi'r(-1 - 2k^2) + \phi(4k^2 - k^4) \\
&= -24ar^4 - 2(24ar + 6b)r^3 + (12ar^2 + 6br + 2c)r^2(1 + 2k^2) \\
&\quad + (4ar^3 + 3br^2 + 2cr + d)r(-1 - 2k^2) + (ar^4 + br^3 + cr^2 + dr + e)(4k^2 - k^4) \\
&= -24ar^4 - 2(24ar^4 + 6br^3) + (12ar^4 + 6br^3 + 2cr^2)(1 + 2k^2) \\
&\quad + (4ar^4 + 3br^3 + 2cr^2 + dr)(-1 - 2k^2) + (ar^4 + br^3 + cr^2 + dr + e)(4k^2 - k^4) \\
&= -Aar^4 - Bbr^3 - Ccr^2 - Ddr - Ee
\end{aligned}
$$

with

$$
\begin{aligned}
-A &= -24 - 48 + 12(1 + 2k^2) + 4(-1 - 2k^2) + (4k^2 - k^4) \\
&= -24 - 48 + 12 + 24k^2 - 4 - 8k^2 + 4k^2 - k^4 \\
&= -64 + 20k^2 - k^4 \\
&= -(k^2 - 4)(k^2 - 16) \\
-B &= -12 + 6(1 + 2k^2) + 3(-1 - 2k^2) + (4k^2 - k^4) \\
&= -12 + 6 + 12k^2 - 3 - 6k^2 + 4k^2 - k^4 \\
&= -(k^4 - 10k^2 + 9) \\
-C &= 2(1 + 2k^2) + 2(-1 - 2k^2) + (4k^2 - k^4) \\
&= 2 + 4k^2 - 2 - 4k^2 + 4k^2 - k^4 \\
&= 4k^2 - k^4 \\
&= -k^2(k^2 - 4) \\
-D &= (-1 - 2k^2) + (4k^2 - k^4) \\
&= -1 - 2k^2 + 4k^2 - k^4 \\
&= -1 + 2k^2 - k^4 \\
&= -(k^2 - 1)(k^2 - 1) \\
-E &= 4k^2 - k^4 \\
&= -k^2(k^2 - 4)
\end{aligned}
$$

so in the end we recover

$$
\begin{aligned}
\rho(r, \theta) &= \frac{1}{k}\sin(k\theta)\frac{Aar^4 + Bbr^3 + Ccr^2 + Ddr + Ee}{r^3} \\
A &= (k^2 - 4)(k^2 - 16) \\
B &= k^4 - 10k^2 + 9 \\
C &= k^2(k^2 - 4) \\
D &= (k^2 - 1)(k^2 - 1) \\
E &= k^2(k^2 - 4)
\end{aligned}
$$

The pressure is obtained as presented before.

Root mean square velocity

$$
\begin{aligned}
v_{rms}^2 &= \iint (v_r^2 + v_\theta^2)r\,dr\,d\theta \\
&= \iint \left[(\frac{\phi\xi'}{r})^2 + (-\phi'\xi)^2 \right] r\,dr\,d\theta \\
&= \iint (\frac{\phi\xi'}{r})^2 r\,dr\,d\theta + \iint (-\phi'\xi)^2 r\,dr\,d\theta \\
&= \underbrace{\int_{R_1}^{R_2} (\frac{\phi}{r})^2 r\,dr}_{I_1} \underbrace{\int_0^{2\pi} (\xi')^2 d\theta}_{I_2} + \underbrace{\int_{R_1}^{R_2} (\phi')^2 r\,dr}_{I_3} \underbrace{\int_0^{2\pi} \xi^2 d\theta}_{I_4}
\end{aligned}
$$

1060

$$
\begin{aligned}
I_1 &= \\
I_2 &= \\
I_3 &= \\
I_4 &=
\end{aligned}
$$

Unfinished! Also compute strain rate tensor, stress, total mass, etc ...

Free slip boundary conditions

what follows needs to be checked!!!

Before postulating the form of $\phi(r)$, let us now turn to the boundary conditions that the flow must fulfill, i.e. free-slip on both boundaries. Two conditions must be met:

- $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ (no flow through the boundaries) which yields $u(r = R_1) = 0$ and $u(r = R_2) = 0$, :

$$
\frac{1}{r} \frac{\partial \Psi}{\partial \theta}(r = R_1, R_2) = 0 \qquad \forall \theta
$$

which gives us the first constraint since $\Psi(r, \theta) = \phi(r)\xi(\theta)$:

$$
\phi(r = R_1) = \phi(r = R_2) = 0
$$

- $(\boldsymbol{\sigma} \cdot \boldsymbol{n}) \times \boldsymbol{n} = \boldsymbol{0}$ (the tangential stress at the boundary is zero) which imposes: $\sigma_{\theta r} = 0$, with

$$
\sigma_{\theta r} = 2\eta \cdot \frac{1}{2} \left(\frac{\partial v}{\partial r} - \frac{v}{r} + \frac{1}{r}\frac{\partial u}{\partial \theta} \right) = \eta \left(\frac{\partial}{\partial r}(-\frac{\partial \Psi}{\partial r}) - \frac{1}{r}(-\frac{\partial \Psi}{\partial r}) + \frac{1}{r}\frac{\partial}{\partial \theta}(\frac{1}{r}\frac{\partial \Psi}{\partial \theta}) \right)
$$

Finally Ψ must fulfill (on the boundaries!):

$$
-\frac{\partial^2 \Psi}{\partial r^2} + \frac{1}{r}\frac{\partial \Psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \Psi}{\partial \theta^2} = 0
$$

$$
-\phi''\xi + \frac{1}{r}\phi'\xi + \frac{1}{r^2}\phi\xi'' = 0
$$

or,

$$
-\phi'' + \frac{1}{r}\phi' - k^2\frac{1}{r^2}\phi = 0
$$

Note that this equation is a so-called Euler Differential Equation[3]. Since we are looking for a solution ϕ such that $\phi(R_1) = \phi(R_2) = 0$ then the 3rd term of the equation above is by definition zero on the boundaries. We have to ensure the following equality on the boundary:

$$
-\phi'' + \frac{1}{r}\phi' = 0 \qquad \text{for} \quad r = R_1, R_2
$$

The solution of this ODE is of the form $\phi(r) = ar^2 + b$ and it becomes evident that it cannot satisfy $\phi(r = R_1) = \phi(r = R_2) = 0$.

Separation of variables leads to solutions which cannot fulfill the free slip boundary conditions

---

[3]http://mathworld.wolfram.com/EulerDifferentialEquation.html

20.20 Rayleigh-Bénard convection for silicon oil

This originates in Jenkins *et al.* (2014) [681].
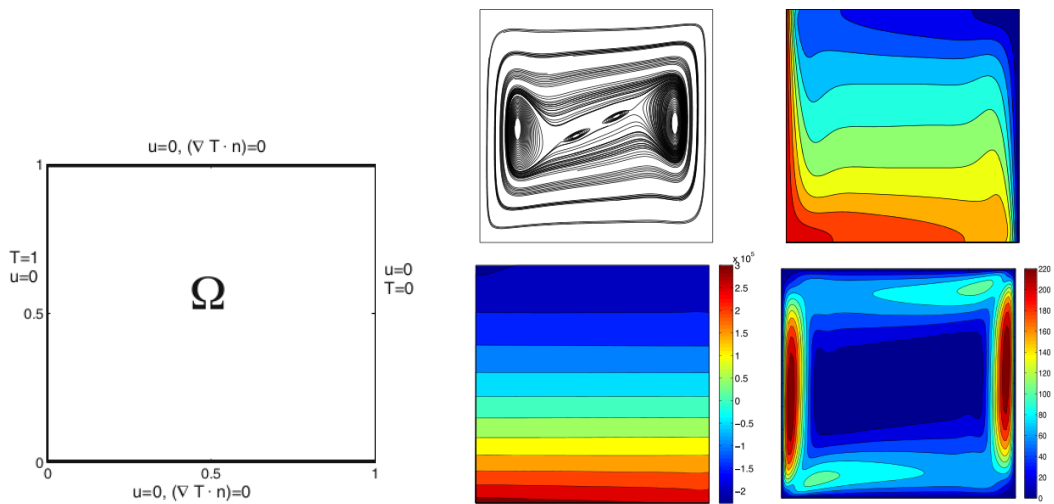
20.21 Rayleigh-Taylor experiment of van Keken *et al.* (1997)

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/benchmark_vaks97

This numerical experiment was first presented in van Keken, King, Schmeling, Christensen, Neumeister, and Doin [1406] (1997). It consists of an isothermal Rayleigh-Taylor instability in a two-dimensional box of size $Lx = 0.9142$ and $L_y = 1$.

Two Newtonian fluids are present in the system: the buoyant layer is placed at the bottom of the box and the interface between both fluids is given by

$$y(x) = 0.2 + 0.02\cos\left(\frac{\pi x}{L_x}\right) \tag{20.97}$$

The bottom fluid is parametrised by its density $\rho_1 = 1000$ and its viscosity η_1, while the layer above is parametrised by $\rho_2 = 1010$ and $\eta_2 = 100$. This experiment is to be carried out for various viscosity constrasts between the two layers, i.e. $\eta_1 = \{1, 10, 100\}$.

No-slip boundary conditions are applied at the bottom and at the top of the box while free-slip boundary conditions are applied on the sides. Gravity is pointing downwards with $|\vec{g}| = 10$.
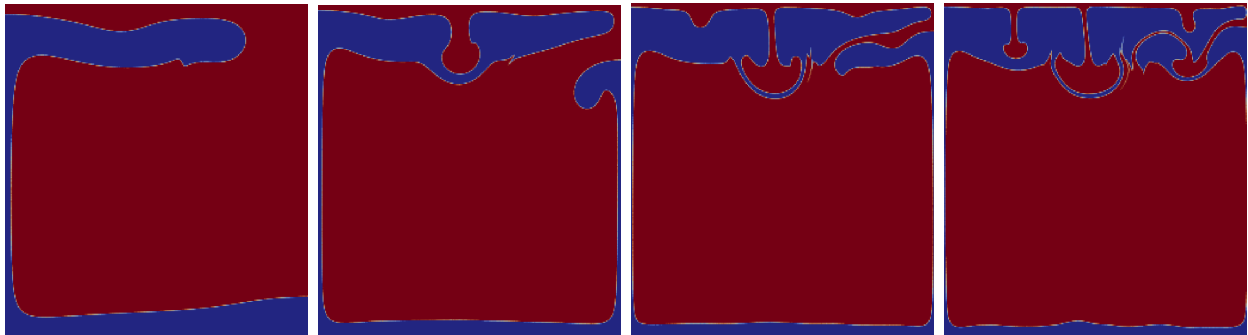
Example of time evolution obtained with the ELEFANT code for a 256x256 grid with the particle-in-cell method.

in images/benchmark_vaks97/

One can measure the following quantities:

- the root mean square velocity ν_{rms} in the domain as a function of time:

$$\nu_{rms} = \sqrt{\frac{1}{L_x L_y} \int \int |\vec{\nu}|^2 dx dy} \tag{20.98}$$

- the maximum (or local maxima) of the ν_{rms} and its (their) corresponding time(s)

- the growth rate of the instability at $t = 0$. From linear stability analysis, the analytical growth rate can be calculated [1105, 1104]: $\gamma_{th} = 0.01094019$, which is valid for an infinitesimal perturbation. For each model run, the growth rate γ is measured by fitting the ν_{rms} and/or (?) the maximum vertical velocity measurements for a short time t.

- the total mass of the system $M(t)$ as a function of time. Since there is no chemical diffusion in the system (pure advection equation) the amount of material in the system is to remain constant, and therefore its mass too.

$$M(t) = \int \int \rho(x, y, t) dx dy \tag{20.99}$$

Given the layout described in the previous paragraph, the exact analytical initial mass M_0 of the system is given by

$$M_0 = 0.9142 \times (0.2 \times 1000 + 0.8 \times 1010) = 921.5136$$

The average density is then

$$< \rho >_0 = \frac{M_0}{L_x L_y} = 1008$$

We will then measure the relative mass error as a function of time

$$\delta M(t) = \frac{M(t) - M_0}{M_0}$$

which is equal to

$$< \delta \rho > (t) = \frac{< \rho > (t) - < \rho >_0}{< \rho >_0}$$

- the length of the interface between the fluids. At startup it is given by

$$\mathcal{L}(0) = \int_0^L \sqrt{1 + (dy/dx)^2} dx$$

with $y(x) = 0.2 + 0.02 \cos(\pi x / L)$. Using WolframAlpha, we find

$$\mathcal{L}(0) = \frac{L}{\pi} \int_0^\pi \sqrt{1 + \left(0.02 \frac{\pi}{L} \sin(\pi x / L)\right)^2} dx \simeq 0.9152786349$$

Instantaneous results . Results obtained with STONE ?? ($Q_2 \times Q_1$ element) and STONE ?? ($P_2^+ \times P_{-1}$ element), both with mesh fitted so as to follow the interface between both fluids.

		$\eta_1 = 100$	$\eta_1 = 10$	$\eta_1 = 1$
$\min(u)$	Stone 25			-4563.5
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\max(u)$	Stone 25			1054.03
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\min(v)$	Stone 25			
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\max(v)$	Stone 25			2070.9
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\max(|v|)$	Stone 25	422.125		4563.67
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
v_{rms}	Stone 25	185.2947		1441.87
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\min(p)$	Stone 25	-5048.16		-5048.7345
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			
$\max(p)$	Stone 25	5033.6947		5032.329
	Stone 93 ($P_2^+ \times P_{-1}$)			
	Aspect			

Code	Grid	Method	Growth Rate γ	t (max vrms)	max (vrms)	publicat
HS	41x41	extrapolated	0.010996	211.1	0.0030958	[1406]
	61x64	extrapolated	0.011109	209.17	0.0031022	
	81x81	extrapolated	0.011177	208.99	0.0030916	
CND	32x32		0.01106	208.4	0.003092	[1406]
	48x48		0.01106	208.5	0.0030943	
SK	80x80		0.01130	215.67	0.00299279	[1406]
	120x120		0.01127	206.38	0.0028922	
	160x160		0.01179	207.84	0.0028970	
PvK	30x30	splines	0.01185	213.38	0.00300	[1406]
	50x50	splines	0.01198	211.81	0.003016	
	80x80	splines	0.01207	210.75	0.003050	
	100x100	splines	0.01211			
	30x30	C1 element	0.01253	210.59	0.003100	
	80x80	C1 element	0.01225	207.05	0.003091	
ISMM	160x160		0.00991	230.1	0.003093	[1255]
ISMM	120x120		0.00998	226.1	0.003133	
MSOU	160x160		0.00993	231.4	0.003085	
MSOU	120x120		0.01020	227.6	0.003134	
FSOU	160x160		0.01111	217.3	0.003118	
FSOU	120x120		0.01159	213.1	0.003151	
	64x64	5	0.01112	206.5	0.003041	[1313]
		15	0.01117	208.8	0.003098	
		40	0.01115	209.9	0.003110	
	128x128	5	0.01113	208.1	0.003079	
		15	0.01110	208.9	0.003097	
		40	0.01109	209.2	0.003102	
	120x132	Level sets	0.01252	211.2	0.00301	[1302]
	67m res.			215.3	0.003106	[256]
	100m res.			215.28	0.003101	
LaCoDe	1808 elts (10754 dofs)		0.01221	215	0.003110	[341]
	7093 elts (2592 dofs)		0.01222	212	0.003080	
	17960 elts (107468 dofs)		0.01222	211	0.003075	

Selected Quantities for the Isoviscous Rayleigh-Taylor problem. HS: FDM, stream function formulation, particles. PvK: FEM, stream function, marker-chain.

SK: FEM, ConMan code, compositional field. CND: spline method, stream function formulation, particles.

List of literature showcasing results of the van Keken *et al.* (1997) [1406] setup:

- de Smet *et al.* (2000) [342]. No table of results, only figures:
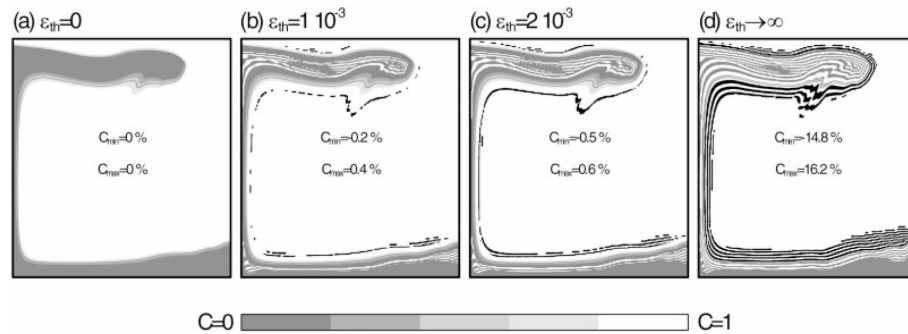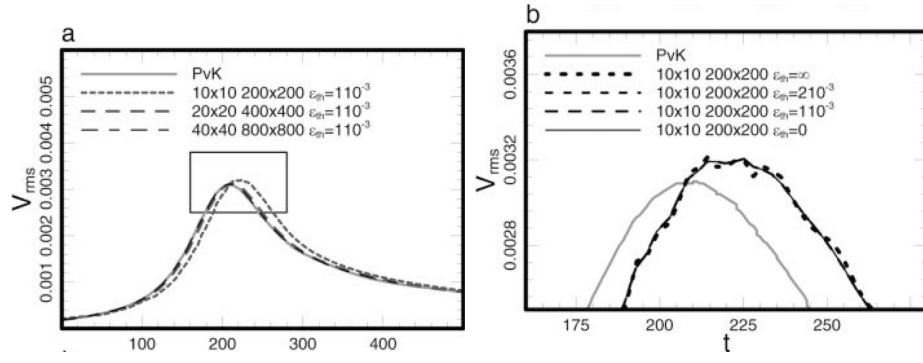
- Soboutia *et al.* (2001) [1255]. Results reported in table above.
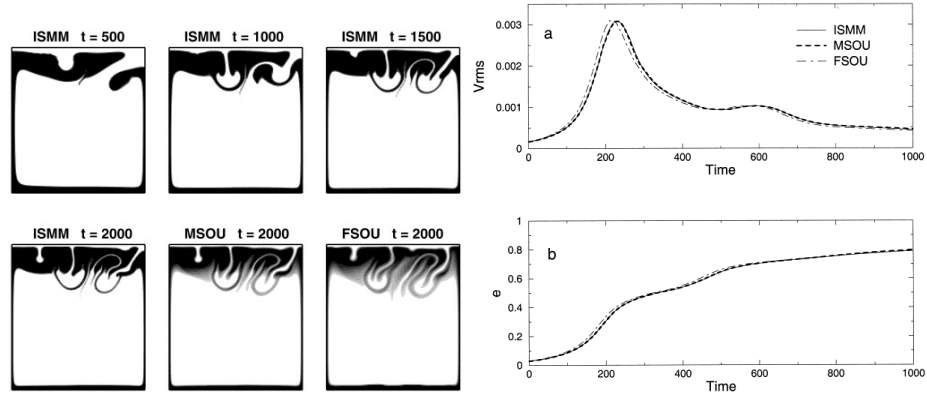
- Babeyko *et al.* (2002) [38]. No table of results, only one figure:
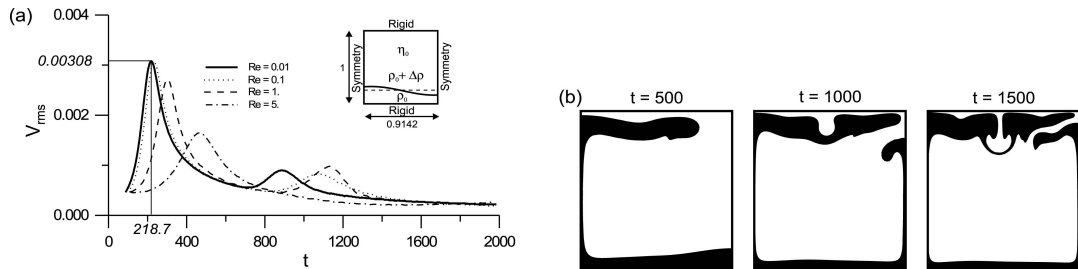
- Tackley & King (2003) [1313]. Performed with grid resolutions of 64x64 or 128x128 and with either 5, 15, or 40 tracers per cell (on average). Results reported in table above.
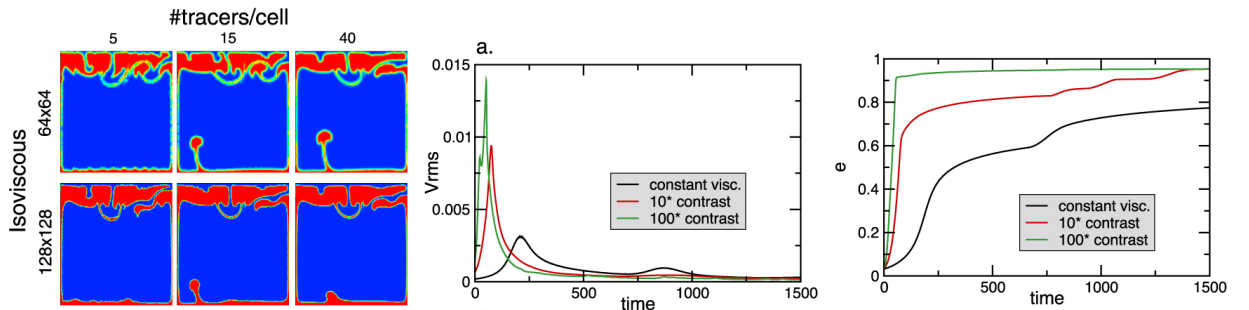
  There are also results for non-isoviscous cases.

- Bourgouin *et al.* (2006) [132]. No table of results, only figures. Additional results for non-isoviscous in the paper.
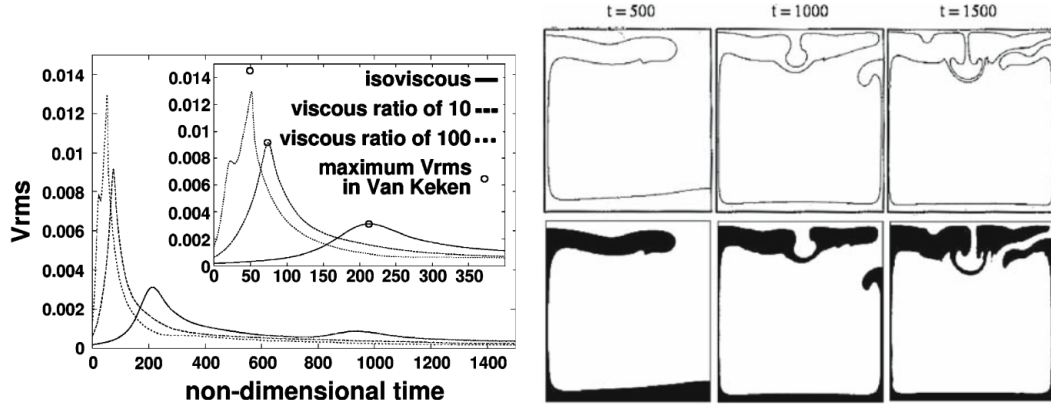
- Quinteros *et al.* (2009) [1095].
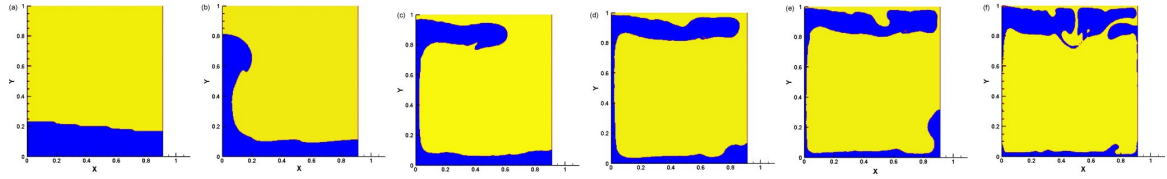
"Different snapshots from the domain evolution that are shown in Fig. 10 were compared with the ones published by van Keken *et al.* (1997). The evolution shown in this chapter and in the van Keken paper are identical for all the compared time steps."

- Samuel & Evonuk (2010) [1177].
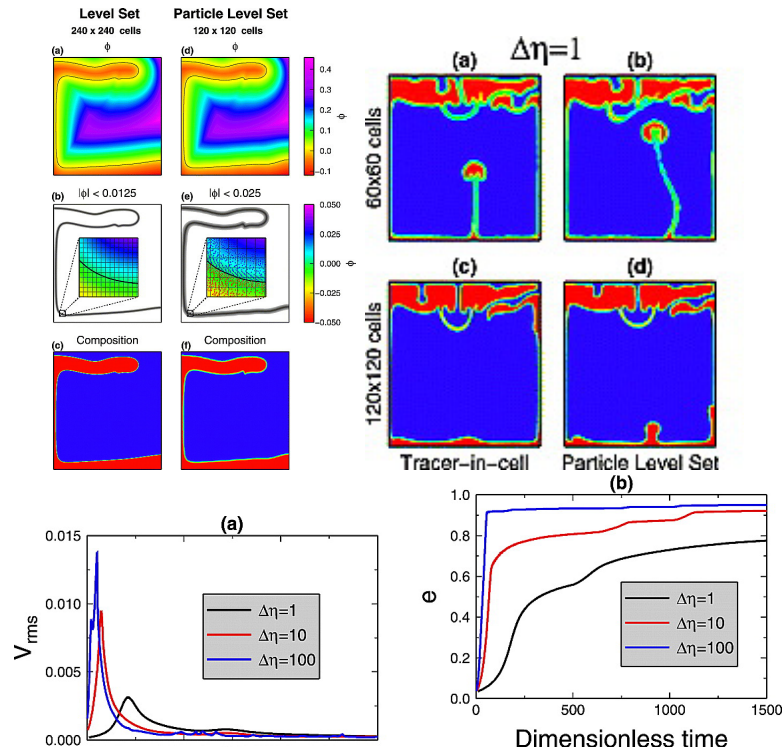
- Suckale *et al.* (2010) [1302].
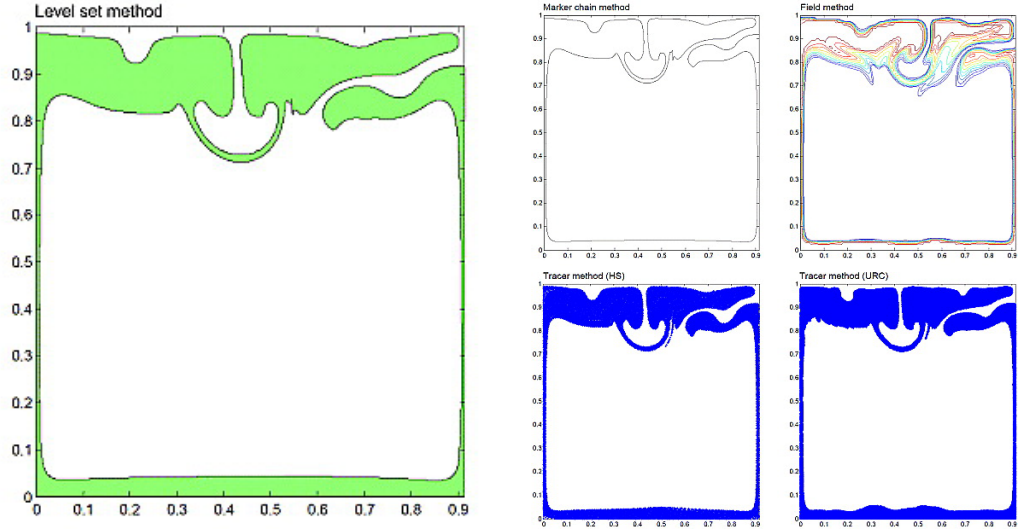
The Rayleigh-Taylor instability at t = 1500 computed by (left) the level set method on a 300x330 grid, compared to the best results

of (right) the four codes compared by van Keken *et al.*
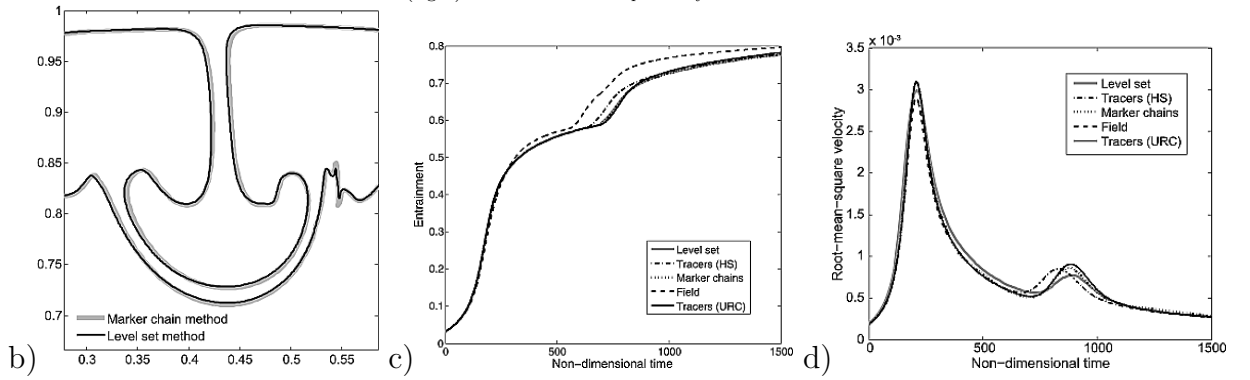
b) Detailed comparison of the level set (thin black line) and the marker chain approach (thick grey line) for the isothermal and isoviscous

Rayleigh-Taylor instability at nondimensional time t = 1500. The plotted interfaces represent a zoom onto the instability descending from the top

downwards in the middle of the box. The two methods yield an almost identical interface. c) Evolution of the entrainment of the buoyant fluid over

time as computed by the five different codes. The level set computation was done on a 160×176 grid. d) Evolution of the root mean square velocity

of the interface over time as computed by the five different codes. The level set computation was done on a 160×176 grid.
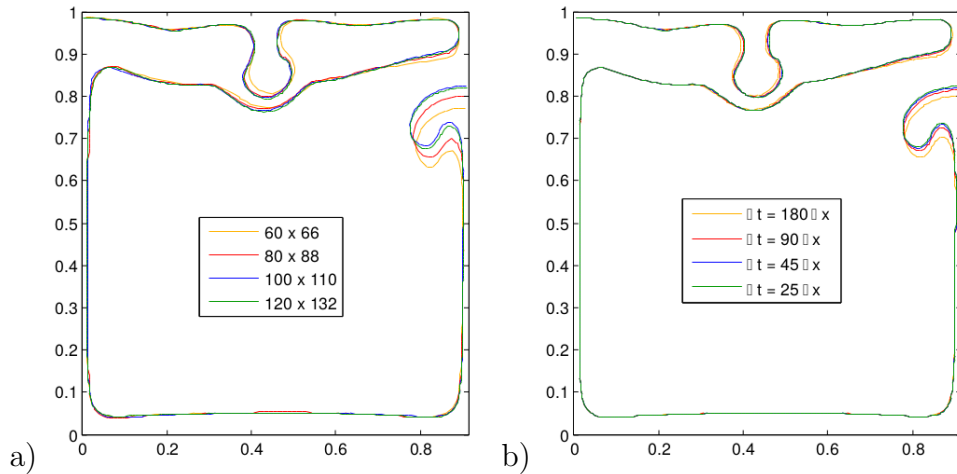
Taken from the supplementary material: a) Convergence test for the isothermal and isoviscous Rayleigh-Taylor instability, benchmark problem 3. A
lack of convergence is easiest to identify during the phases of rapid rise of an instability. We illustrate this for the rise of the secondary instability on
the right side of the box at time t=1000 and four different grid sizes: 60x66, 80x88, 100x110, and 120x132. We observe convergence for grid sizes
above 100x110.

b) Convergence test for the isothermal and isoviscous Rayleigh-Taylor instability, benchmark problem 3. We illustrate this convergence test for the

rise of the secondary instability at time $t = 1000$. The four interfaces were computed based on the time steps: $\Delta t = 180\Delta x$, $\Delta t = 90\Delta x$, $\Delta t = 45\Delta x$,

and $\Delta t = 25\Delta x$. We observe convergence for time steps $\Delta t \leq 25\Delta x$.
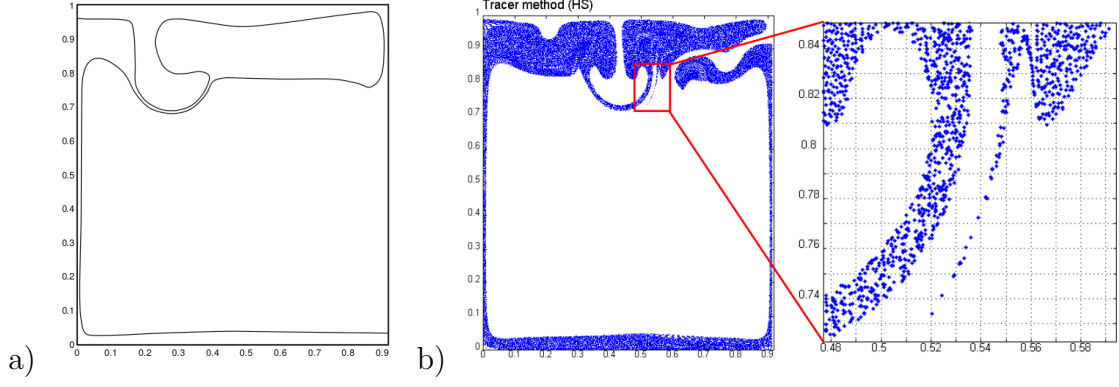
a) The isothermal Rayleigh-Taylor instability with viscosity contrast 10 at non-dimensional time t=500. The computation was done with a grid resolution of 250×275. b) The Rayleigh-Taylor instability as computed by the HS-tracer method at time t=1500. The equations of motion for this simulation were solved on an 81×81 grid. The right panel is a zoom onto the peak located left of the descending instability. Each blue dot represents one particle and the grid represents a rough estimate of the scale at which the flow field is approximated correctly.

- Leng & Zhong (2011) [815].
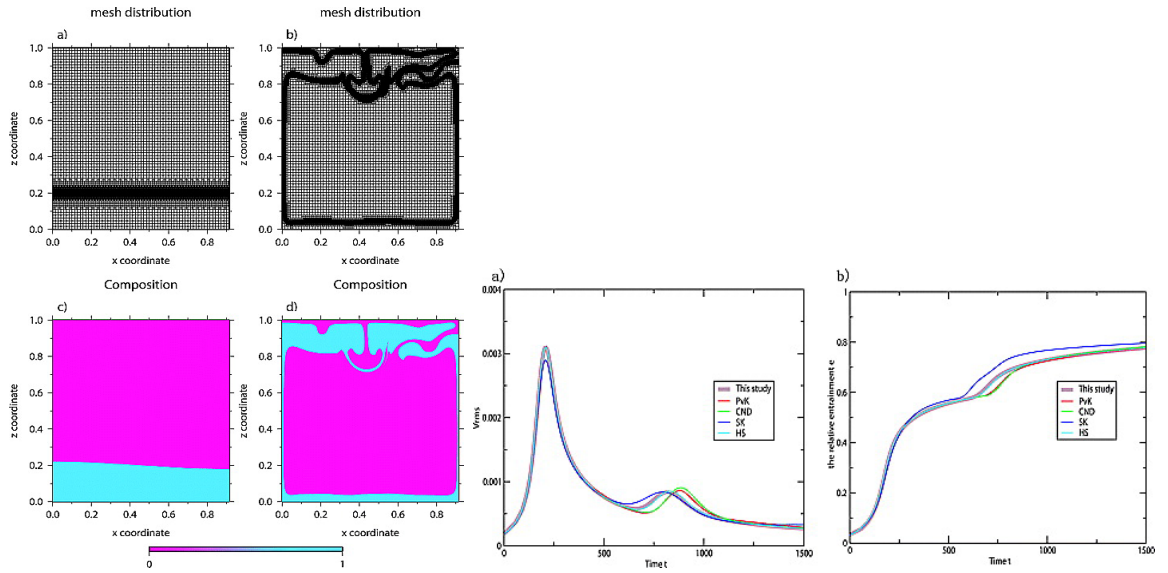
Left: The mesh distribution and chemical composition for case RT1 at two different times: (a and c) t = 0 and (b and d) t = 1500. Right: a) The root mean square velocity v_{rms} and (b) the relative entrainment of the buoyant material, with time for case RT1. The corresponding benchmark results from van Keken *et al.* [1997] are also plotted.

- Vynnytska *et al.* (2013) [1432]. Results only for 100 viscosity contrast

- Choi *et al.* (2013) [256].
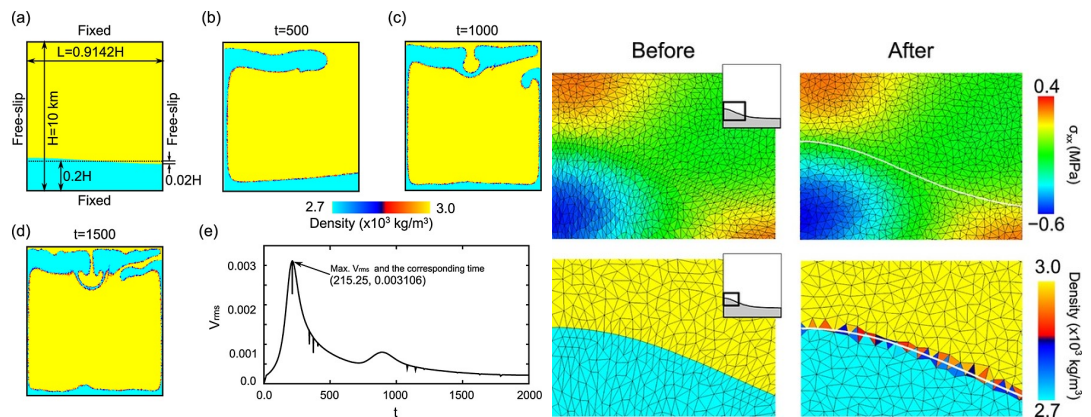
Left: Rayleigh-Taylor instability. (a) Model setup. Snapshots of the density at dimensionless time of (b) 500, (c) 1000, and (d) 1500. (e) Plot of v_{rms} versus dimensionless time, t. The resolution is about 0.6km. Right: σ_{xx} and density fields before and after the first remeshing with about 1km resolution. The white lines in the "After" images denote the original phase boundary before remeshing. The thick-lined box in the inset shows the location of the zoomed-in part of the domain.

- Fuchs & Schmeling (2013) [447].
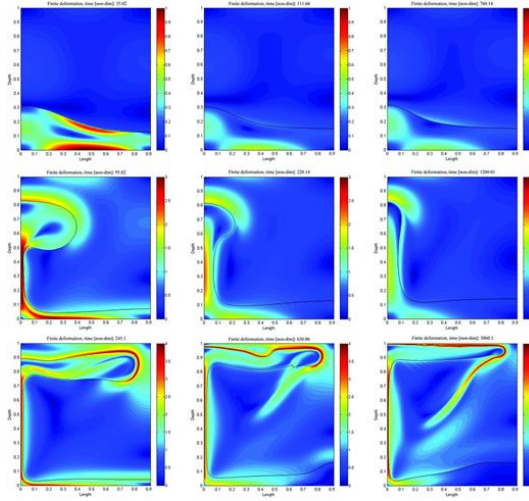
Finite deformation field for different stages of the diapirism with three different viscosity ratios and an initial non-dimensional thickness of the buoyant layer $h_2 = 0.2$. Left column: viscosity ratio $m = 0.1$. Middle column: $m = 1$. Right column: $m = 10$. Top row: pillow stage. Middle row: rising stage. Lower row: final stage.

- de Montserrat *et al.* (2019) [341].
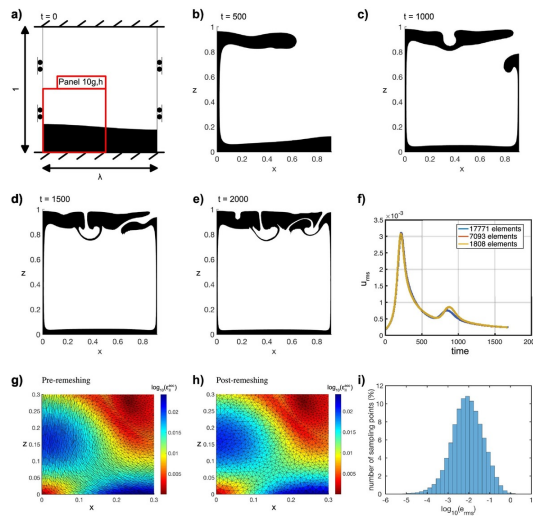
a-e) Temporal evolution of the Rayleigh-Taylor instability. f) Evolution of v_{rms}. Remeshing of the domain is necessary when the mesh becomes highly distorted. Note that the red lines overlap with the blue line. g) Second invariant of the accumulated strain in a mesh with heavily distorted elements, and h) interpolated into a new high-quality mesh. i) Histogram showing the logarithm of the error between the accumulated square root of second invariant of the strain rate, pre- and post-remeshing.

- Robey & Puckett (2019) [1150], Robey (2019) [1149] (PhD thesis).
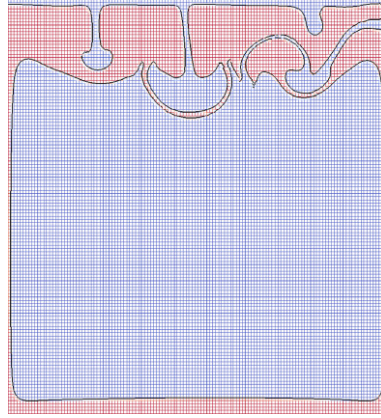
Computed solution of the van Keken isoviscous Rayleigh-Taylor problem at time $t = 2000$ on a uniform grid of 128x128 cells.

- Louis-Napoleon *et al.* (2020) [857].

  Raw data available in `./images/benchmark_vaks97/louis_napoleon_etal`.
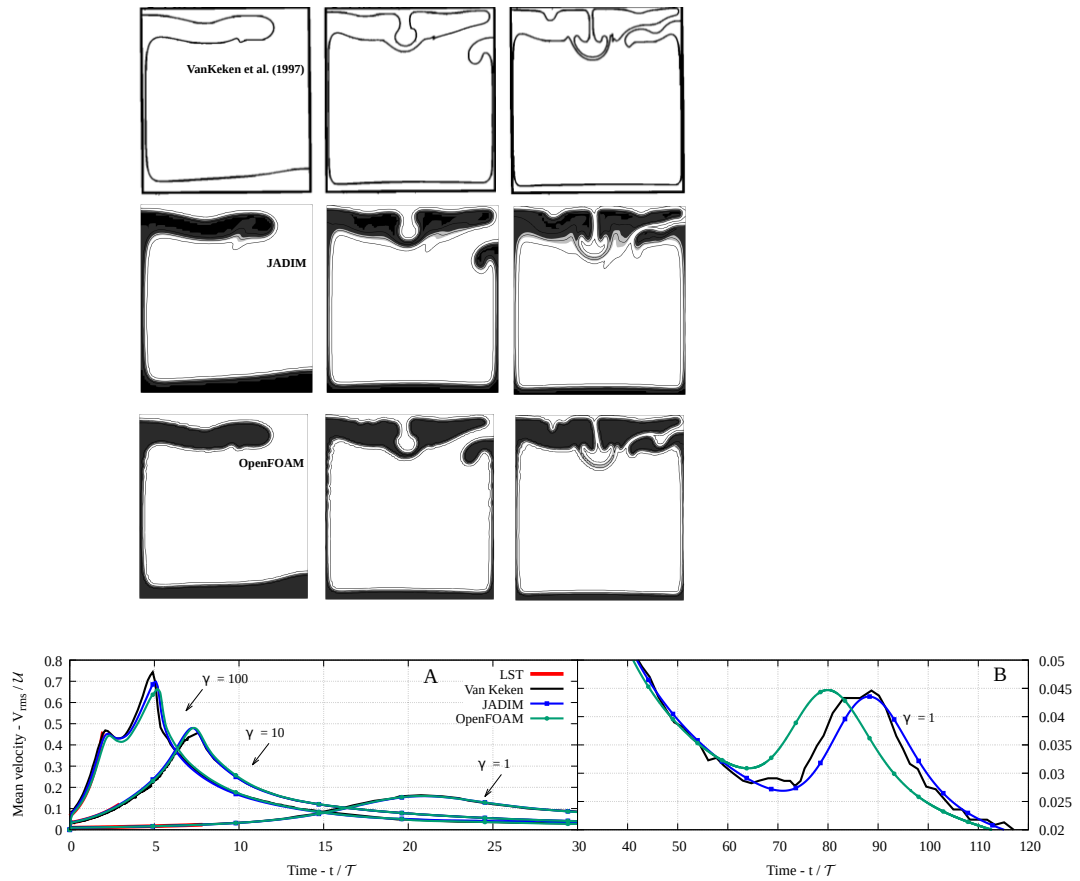
- Maierova (2012) [871] (phd thesis)

Time 500,1000, 2000.
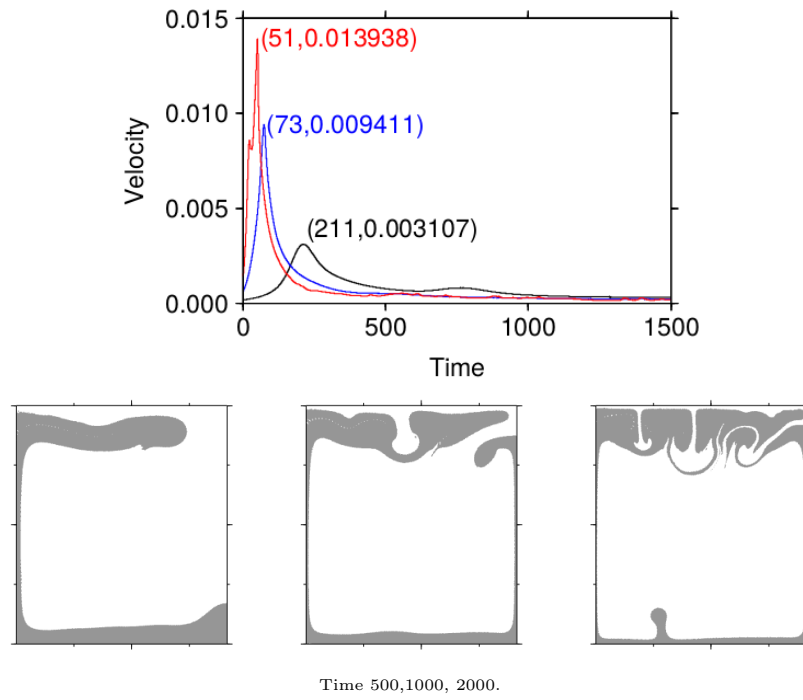
- Schuh-Senlis *et al.* (2020) [1221]. The setup in the paper is inspired by [1406] but results are therefore not consistent with those of [1406].
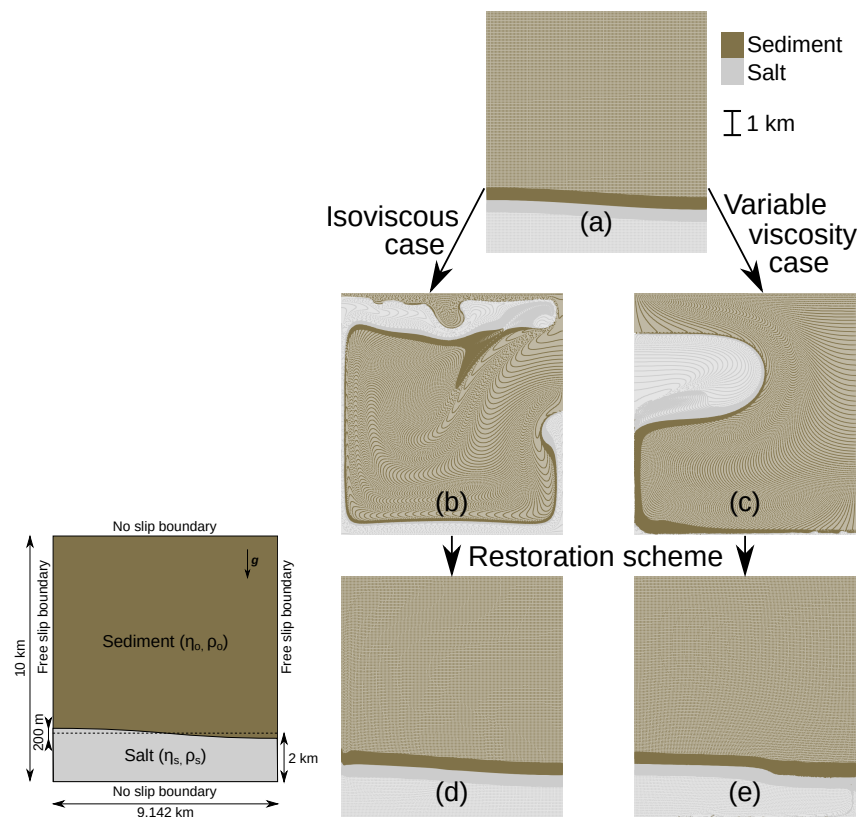
- MVEP2 code, curtesy of Marcel Thielmann
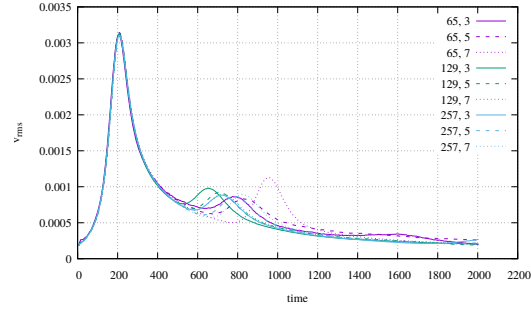
- Burcet, Oliveira, Afonso, and Zlotnik [189] (2024).
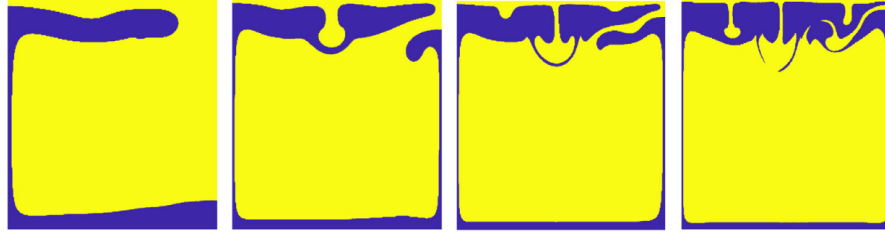
Table 2
Comparison of measurable results in the Rayleigh-Taylor benchmark 4.3.

Method	φ	u_{rms}^{max}	$t(u_{rms}^{max})$	Ref.
FCFV (this study)	0.011093	0.003016	209.35	-
ALE	0.01222	0.003075	211	[58]
Level sets	0.01252	0.00301	211.2	[59]
Tracers (HS)	0.011177	0.003092	208.99	[56]
Tracers (CND)	0.01106	0.003094	208.5	[56]
Field (SK)	0.01179	0.002897	207.84	[56]
Marker Chains (PvK)	0.01225	0.003091	207.05	[56]

- Logg *et al.* (2012) [852]: only Entrainment of a Dense Layer by Thermal convection.

- ASPECT manual [48].
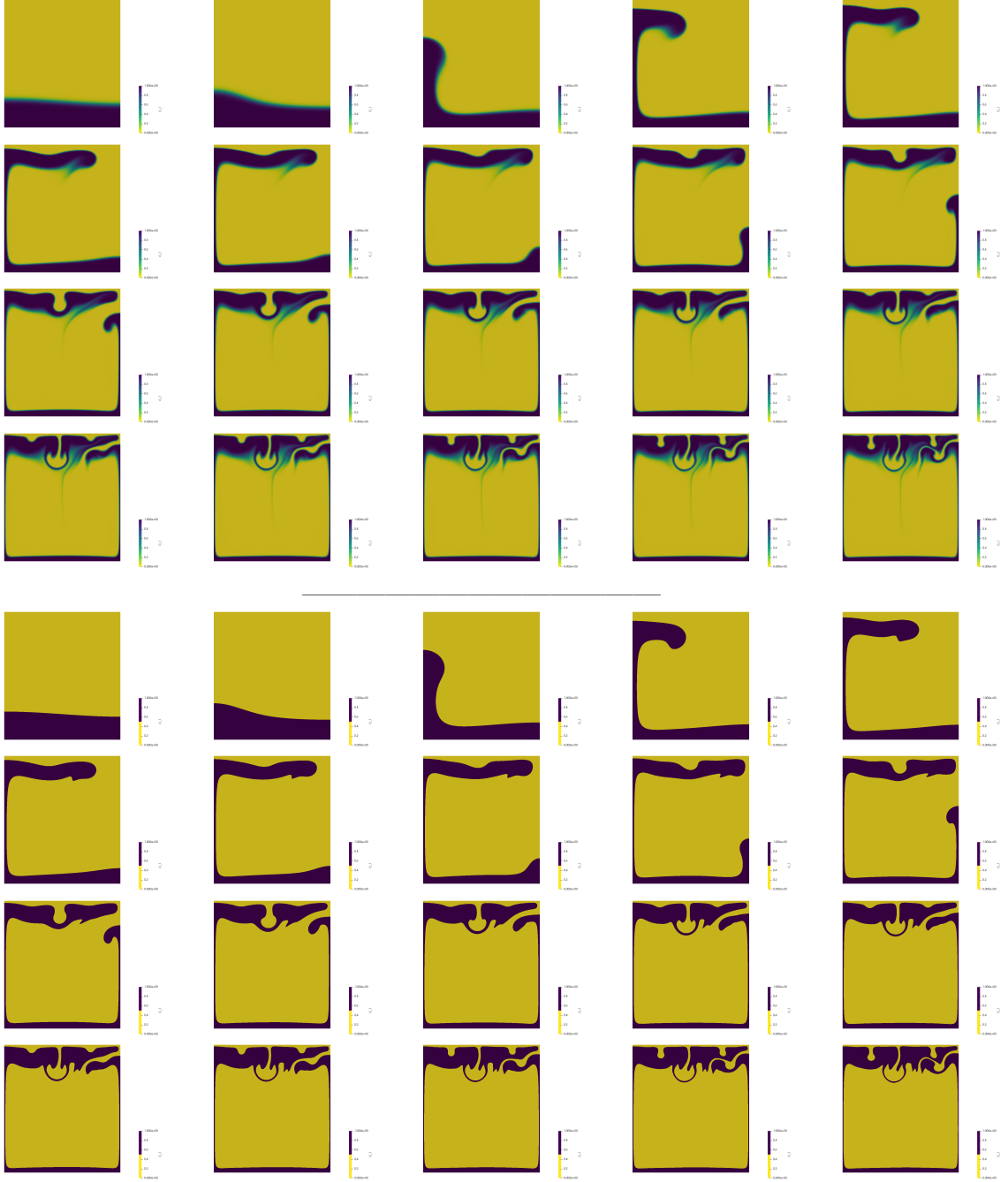
Time evolution of the system, smooth compositional field, CFL=0.25, mesh refinement level 8. Top block is with 256 colors, bottom block is the same data but only using two colors. From t=0 until t=2000, every 100s.

When compositional fields are used the solution has been shown (see ASPECT manual) to be very sensitive to the resolution. In an attempt to remedy this issue, an approach was devised, which consists in replacing the original discontinuous initial condition with a smoothed out version. The function in the input file

```
set Function expression = if( (z>0.2+0.02*cos(pi*x/0.9142)) , 0 , 1 )
```

is then replaced by

```
set Function expression = 0.5*(1+tanh((0.2+0.02*cos(pi*x/0.9142)-z)/0.02))
```

The last number on the line is a parameter which controls the 'thickness' of the interface. When it becomes small we recover a discontinuous implementation (this very much depends on the mesh size).
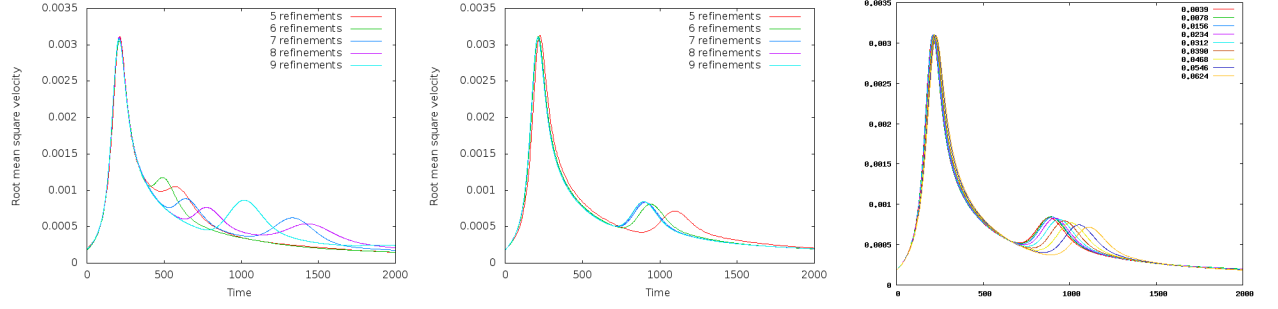
Taken from the ASPECT manual. Root mean square measurements with discontinuous (left) and smoothed, continuous (middle) initial conditions for the compositional field: 5 global refinements correspond to a 32×32 mesh, 9 refinements to a 512×512 mesh. Right: influence of smoothing parameter (level 7).
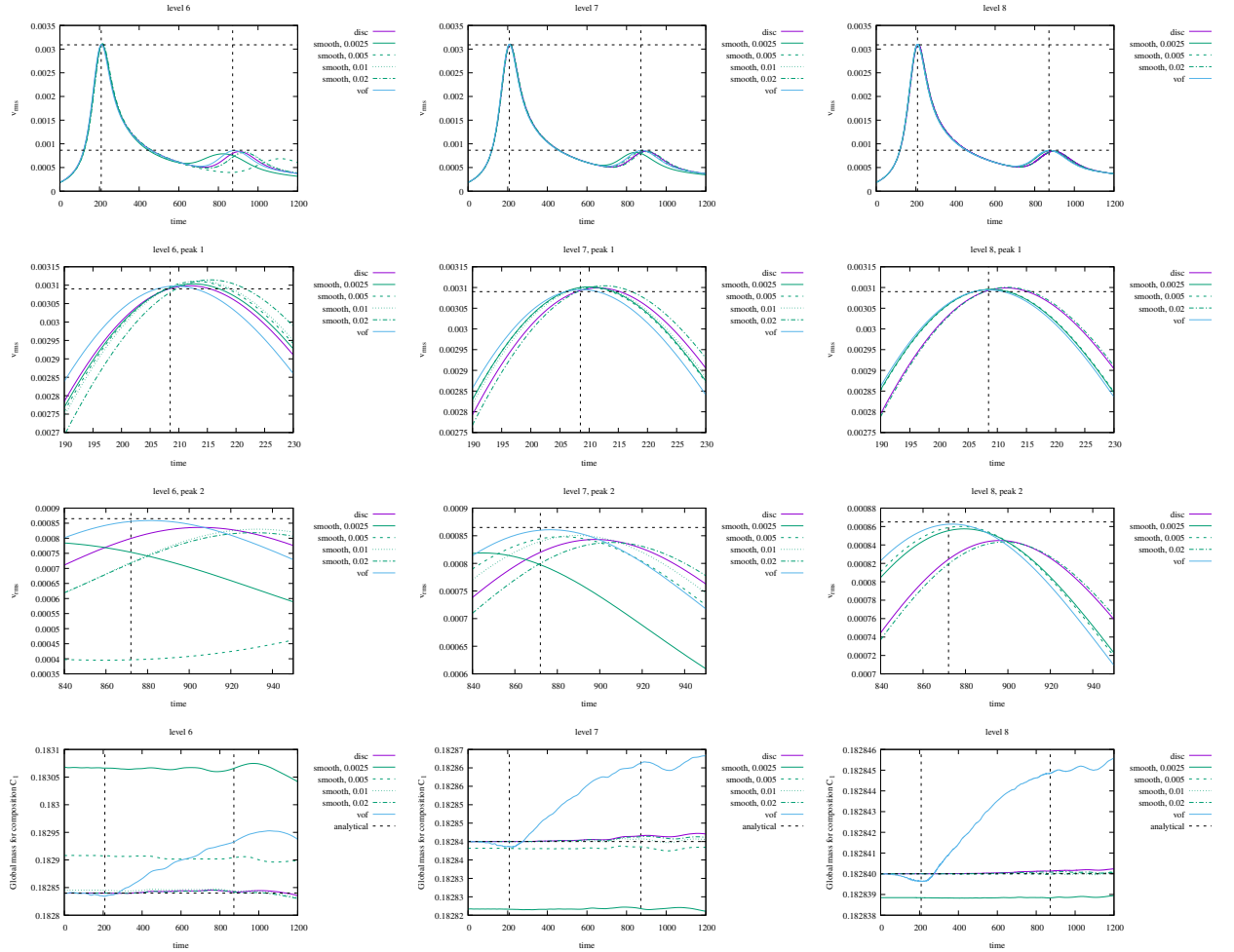
Results obtained for both approaches and vof, on various meshes and for various smoothing parameters. The dashed lines indicate the STONE 95 results.
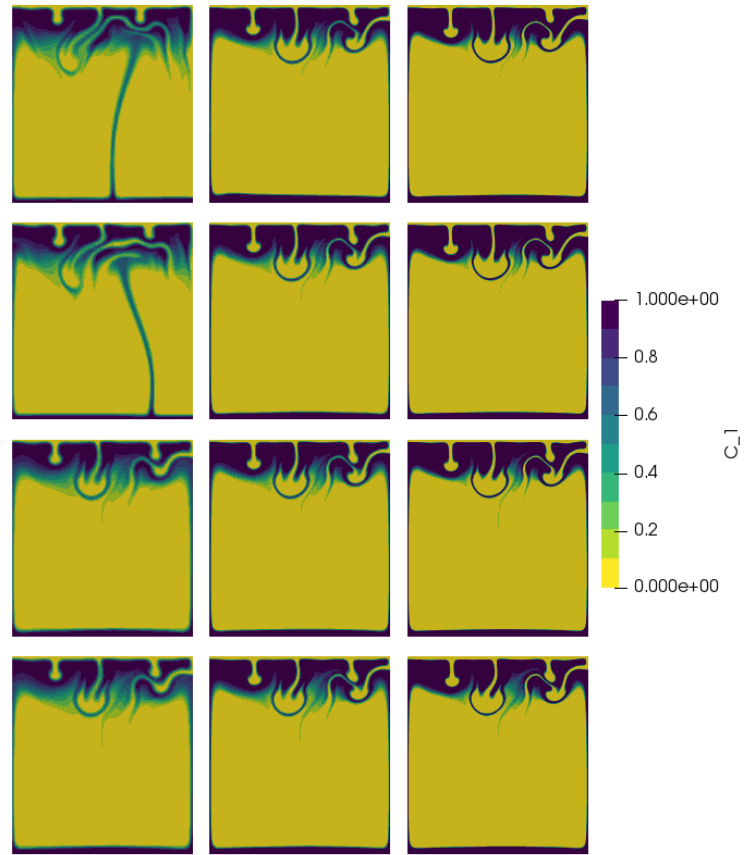
Results for smoothed approach. Left to right: level 6,7,8. Top to bottom: smoothing parameter 0.0025, 0.005, 0.01, 0.02. We see that level6 results in combination with a small smoothing parameter yield a very different outcome.
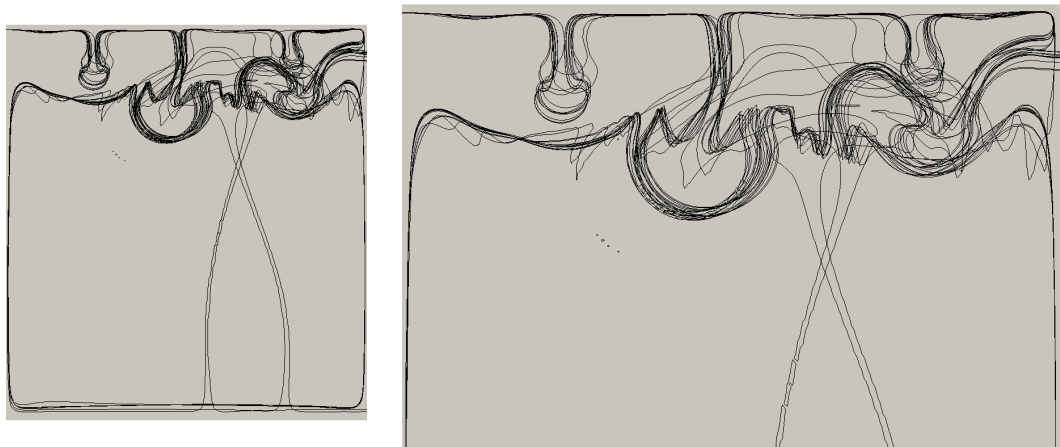
0.5 compositional field C_1 isocontours for all 12 simulations.
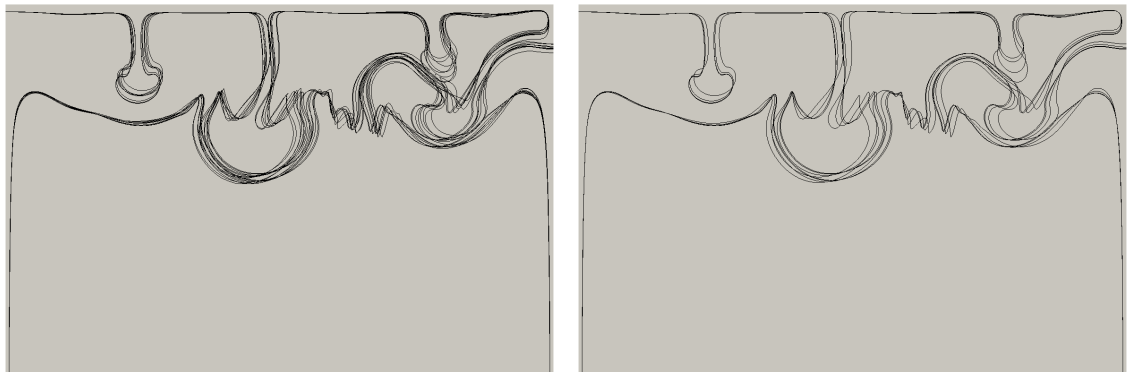
0.5 compositional field C_1 isocontours for levels 7 and 8 (left), and level 8 (right)

The code can also rely on the VOF method to solve the advection equation for the computational fields:
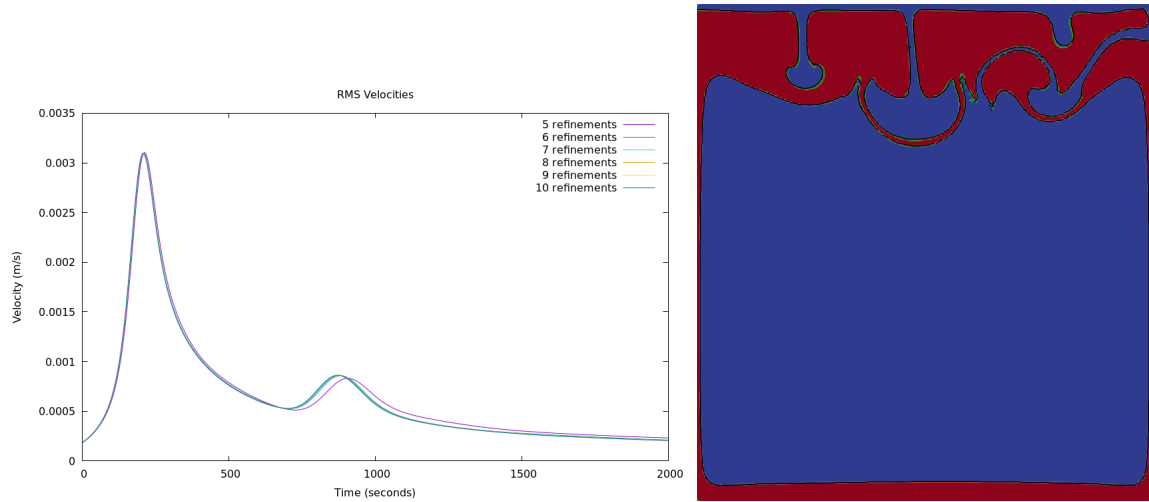
Left: Evolution of the root mean square velocity as a function of time for computations of the van Keken problem made with the VOF interface tracking algorithm with five different global mesh refinements (from a 32×32 mesh to a 1024×1024 mesh. Right: The results of two computations of the van Keken problem made with the VOF interface tracking algorithm overlaid upon each other at $t_{end} = 2000$. This visualization shows the reconstructed boundary between the two materials at the final time t_{end} as computed on a uniform grid with 7 and 8 levels of refinement. The boundaries between the materials are displayed as contours of the fields $\tilde{\psi}^7(t_{end})$ (black) and $\tilde{\psi}^8(t_{end})$ (bright green), which are generated by the visualization postprocessor. The contours for the reconstructed material boundaries are superimposed on a color gradient visualization of the material composition for the computation with 8 levels of refinement in order to make the regions with each fluid type more evident.

- Mulyokova phd thesis [973]
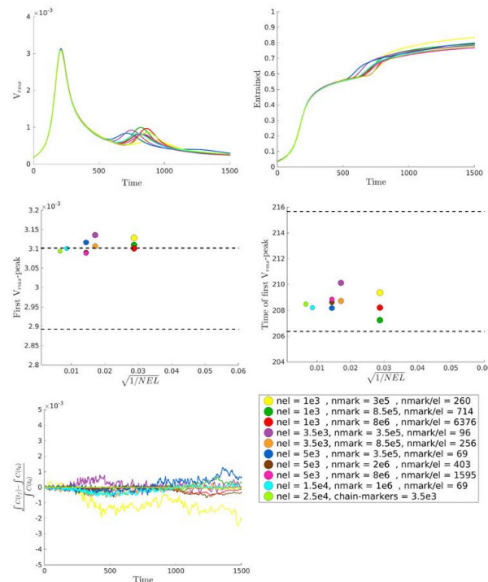
Figure 4.19: *Benchmark-quantities for the isoviscous Rayleigh-Taylor instability set-up, obtained with the tracer-ratio method at different resolutions (according to legend) and marker-chain method with a conforming grid. In the legend, 'nel' means number of mechanical elements, 'nmark' number of markers in the tracer-ratio method, and 'nchain' number of markers in the marker-chain method. Black dashed lines in plots showing values and points in time of the first v_{rms}-peak (middle row) indicate the range of values for the respective quantities presented in Van Keken et al. (1997).*
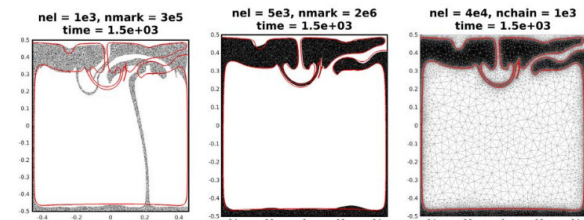
Figure 4.18: *Comparison of the results for the Rayleigh-Taylor benchmark obtained with the tracer-ratio methods at two different resolutions (first two figures), and those obtained with the marker-chain method with a conforming grid, at time $t = 1500$. For the latter, the mesh is shown with gray lines delineating element-edges. The red line delineates the marker-chain, and is plotted for comparison in all three figures. Resolutions are given in the titles of each figure, where 'nel' means number of mechanical elements, 'nmark' number of markers in the tracer-ratio method, and 'nchain' number of markers in the marker-chain method.*

- Maljaars, Richardson, and Sime [877] (2021).

1078

(a) $t = 500$ (b) $t = 1000$ (c) $t = 1500$
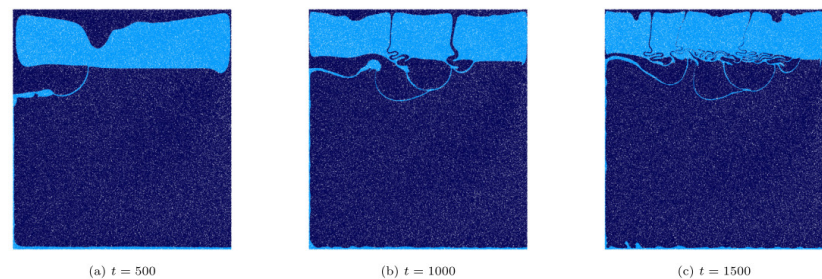
Computed functionals from the Rayleigh–Taylor instability simulation and comparison with other works.

Code	Grid	Method	γ	$t\,(\max u_{\mathrm{rms}})$	$\max u_{\mathrm{rms}}$
van Keken et al. [12]	100×100	Splines / marker chain	0.1024	51.12	0.01385
	80×80	$C1$–element / marker chain	0.1025	51.23	0.01448
Vynnytska et al. [50]	40×40	Taylor–Hood / DG	0.0976	57.21	0.01140
	80×80	Taylor–Hood / DG	0.1018	52.11	0.01444
	160×160	Taylor–Hood / DG	0.1039	51.55	0.01458
This work	40×40	HDG / PDE-constrained projection	0.08707	54.81	0.01208
	80×80	HDG / PDE-constrained projection	0.09161	50.68	0.01428
	160×160	HDG / PDE-constrained projection	0.09677	50.80	0.01436

Results obtained with $\eta^{\star} = 100$.

- Sime, Maljaars, Wilson, and Keken [1246] (2021)
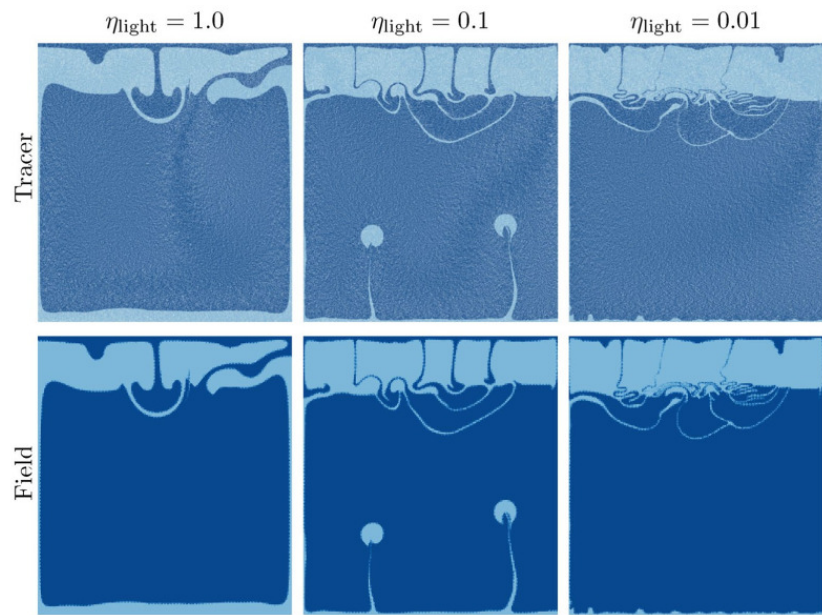
Publication	$\eta^\star$	Stokes method	Transport method	vrms	png	max time	peak 1 time	peak 1 vrms	peak 2 time	peak 2 vrms	growth rate rate	observation
van Keken *et al.* [1406]	1			√	√	2000	211.1	0.0030958			0.010996	HS 41x41
	1						209.17	0.0031022			0.011109	HS 61x61
	1						208.99	0.0030916			0.011177	HS 81x81
	1						208.4	0.003092			0.01106	CND 32x32
	1						208.5	0.0030943			0.01106	CND 48x48
	1						215.67	0.00299279			0.01130	SK 80x80
	1						206.38	0.0028922			0.01127	SK 120x120
	1						207.84	0.0028970			0.01179	SK 160x160
	1										0.01220	SK 160x160
	1						213.38	0.00300			0.01185	PvK 30x30 splines
	1						211.81	0.003016			0.01198	PvK 50x50 splines
	1						210.75	0.003050			0.01207	PvK 80x80 splines
	1										0.01211	PvK 100x100 splines
	1						210.59	0.003100			0.01253	PvK 30x30 C1
	1						207.05	0.003091			0.01225	PvK 80x80 C1
de Smet *et al.* [342]												
Sobouti *et al.* [1255]												
Babeyko *et al.* [38]												
Tackley & King [1313]												
Bourgouin *et al.* [132]												
Quinteros *et al.* [1095]												
Samuel & Evonuk [1177]												
Suckale *et al.* [1302]	1	FDM	level set	√	√	1500	211.2	0.00301	?	?	0.01252	120x132
Suckale *et al.* [1302]	10	FDM	level set	√	√	500	?	?	?	?	0.04809	250x275
Leng & Zhong [815]	1	FEM	tracers	√	√	1500	?	?	?	?	?	AMR(6+3)
Maierova PhD thesis [871]	1			Y	Y	2000	211	0.003107				
	0.1			Y	N	2000	73	0.009411				
	100			Y	N	2000	51	0.013938				
Vynnytska *et al.* [1432]												
Choi *et al.* [256]												
Fuchs & Schmeling [447]												
Mulyukova PhD thesis[973]												
de Montserrat *et al.* [341]												
Robey & Puckett [1150]												
Robey PhD thesis [1149]												
Trim *et al.* (2021) [1372]												
Louis-Napoleon *et al.* [857]	1	FEM	VOF & CF									JADIM, OpenFOAM & Aspect
Aspect manual [48]	1	FEM	CF, PIC	√	√							

20.22 (Instantaneous) Sinking block (2D)

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/sinking_block

The domain is a unit square. Fluids are such that $\rho_1 = 1$, $\eta_1 = 1$ and $\rho_2 = 1.01$, $\eta_2 = 1000$. Boundary conditions are either free slip or no slip on all sides. Pressure is normalised so that the volume average is zero. Gravity points downwards with $|\vec{g}| = 1$. Profile measurements are carried out on the dashed line.
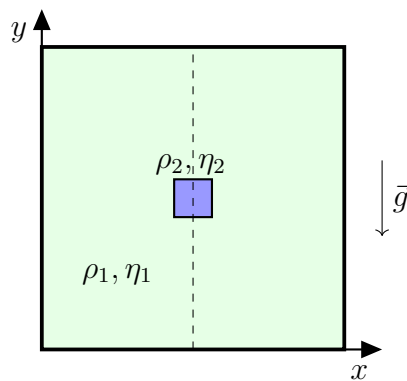
When using ASPECT , it is good to remember that a compositional field is used, which 'lives' on the nodes of the FE grid. Part of the input file is shown here:

```
subsection Compositional fields
  set Number of fields = 1
end

subsection Initial composition model
  set Model name = function
  subsection Function
    set Variable names      = x,y
    set Function constants  = p=0.5
    set Function expression = if(abs(x-p)<0.0625 && abs(y-p)<0.0625 , 1, 0)
  end
end

subsection Material model
  subsection Simple model
    set Density differential for compositional field 1 = 0.01
    set Composition viscosity prefactor = 1000
  end
end
```

The value of the composition (and therefore the density and viscosity values) on a quadrature point is obtained via interpolation and averaging, which is different than the STONE codes where the density and viscosity are elemental quantities.

Free-slip boundary conditions .
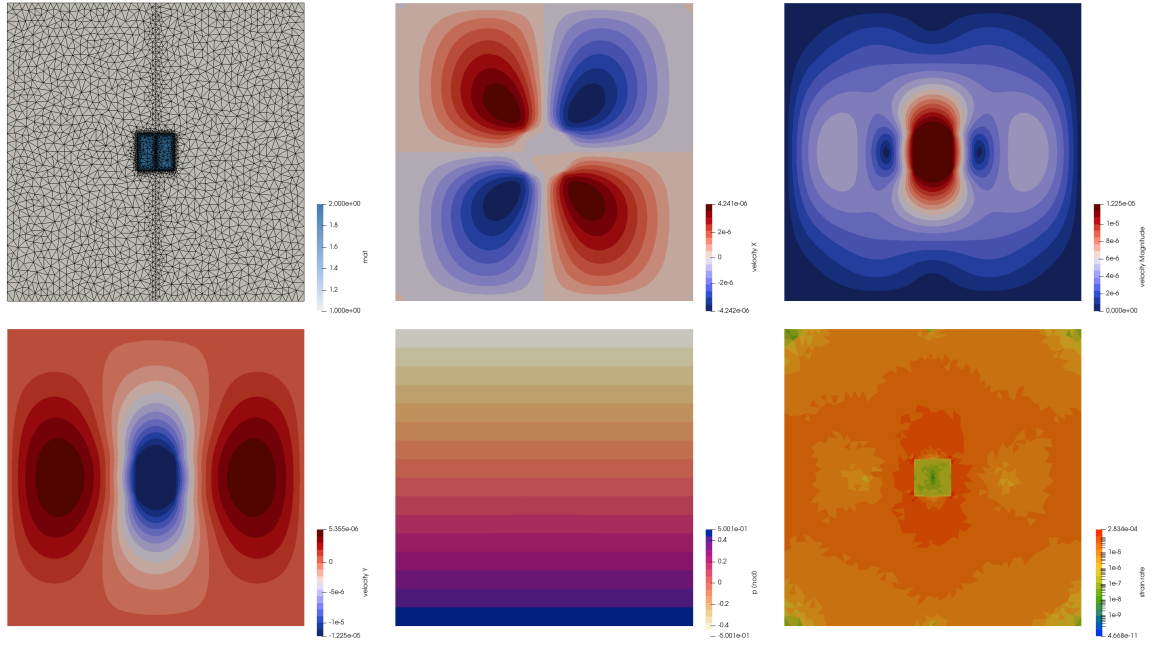
Results obtained with Stone 93.
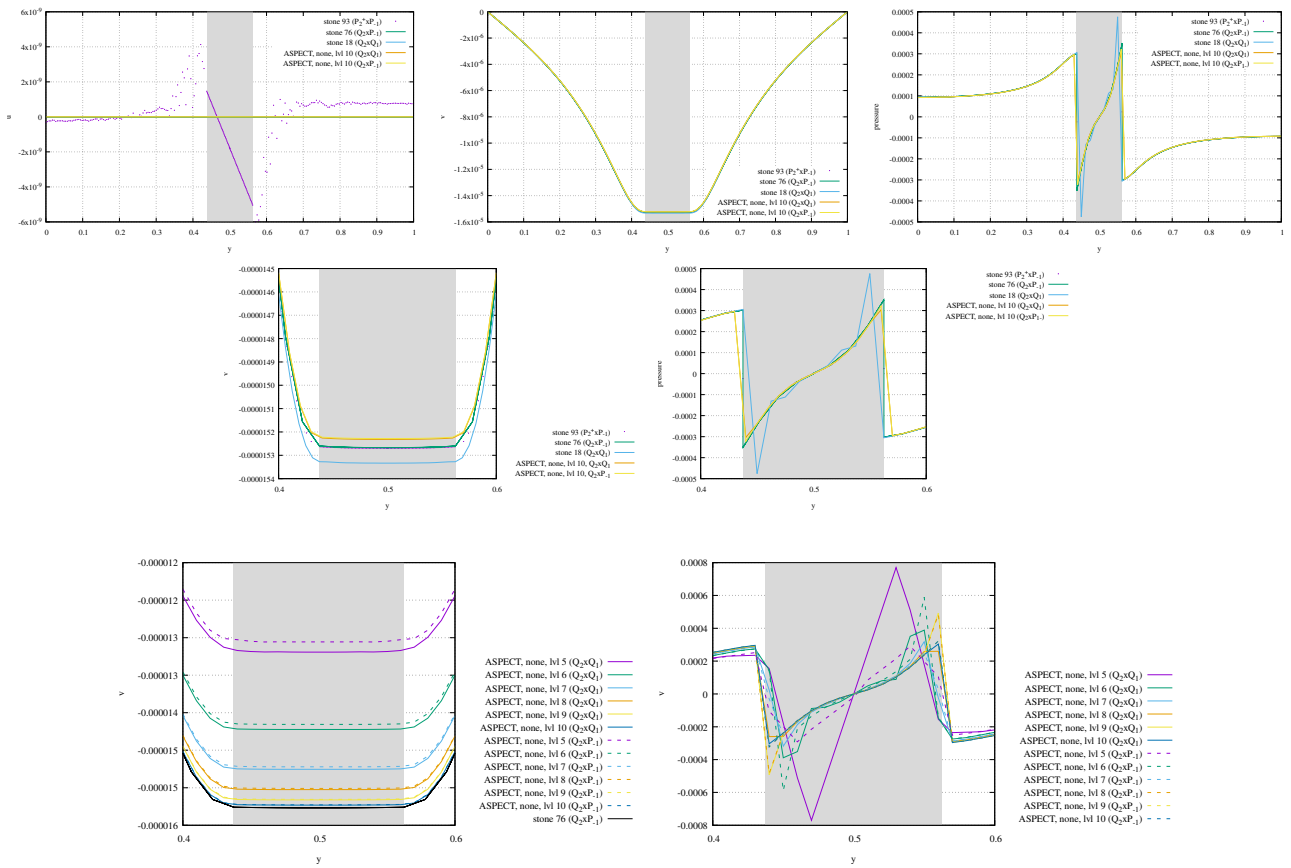
ASPECT results with various global mesh refinement. No averaging
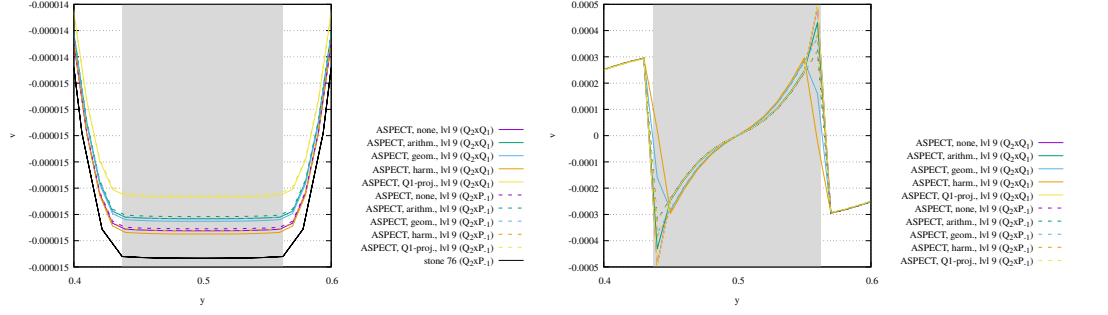
ASPECT results with various averagings. level 9.

Using error extrapolation (see Section 16.11), one can compute an estimate of the resolution independent value of the vrms of maximum velocity for example:
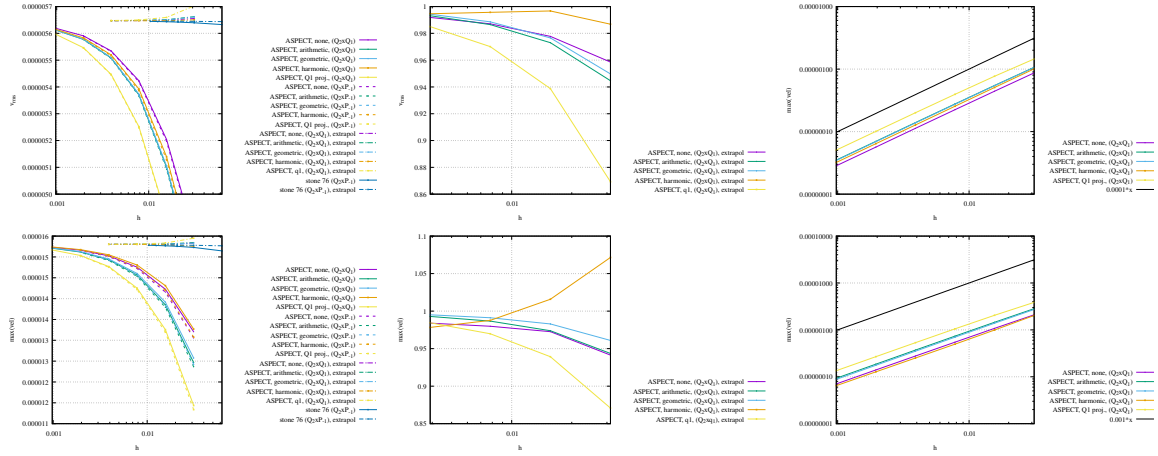
We find that the convergence rates are near unity.

No-slip boundary conditions .
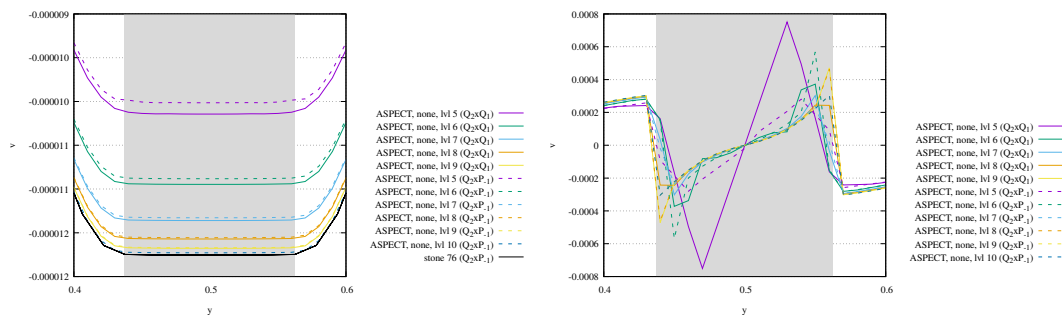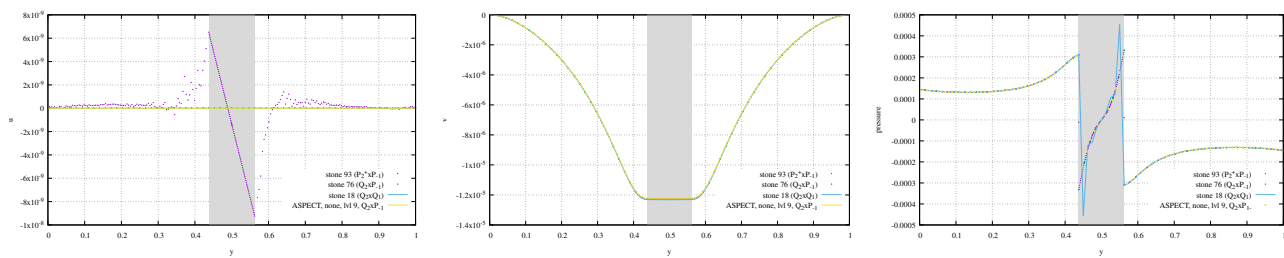
ASPECT results with various global mesh refinement. No averaging
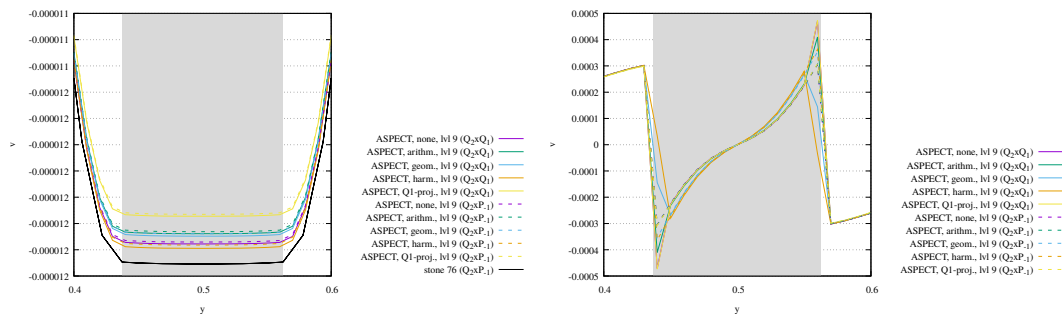
ASPECT results with various averagings. level 9.
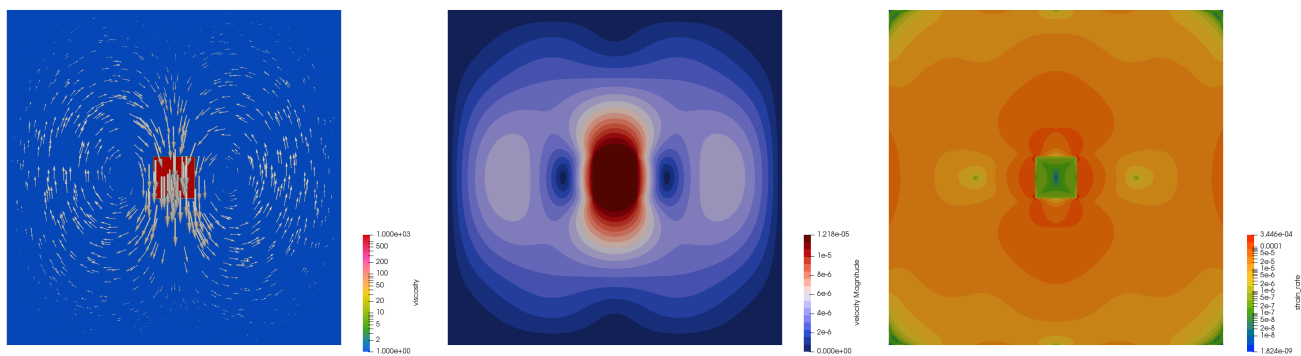
Obtained with ASPECT, level 9.

TODO: finish analysis

20.23 (Instantaneous) Stokes sphere (3D)

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/stokes_sphere3D

This is a simple experiment without an analytical solution. The idea here is simple: to design a small number of Stokes sphere-related experiments and provide for them (very) high-resolution results obtained with various codes so as to turn these into benchmarks. The domain is chosen to be a unit cube. Gravity is such that $\vec{g} = (0, 0, -g)$ with $g = 1$. The sphere is in the middle of the domain and has a radius $R = 0.123456798$. The fluid has a density $\rho_f = 1$ and viscosity $\eta_f = 1$. The sphere has a density $\rho_s = \rho_f + \delta\rho$ and a viscosity $\eta_s = 10^m \eta_f$. Default values for $\delta\rho$ and m are set to 0.01 and 3 respectively.

Concerning boundary conditions, we distinguish three cases:

- FS: free slip boundary conditions are imposed on all 6 sides;

- NS: no slip boundary conditions are imposed on all 6 sides;

- OT: free slip boundary conditions are prescribed on the sides and bottom, but the top surface is open.

- BO: both top and bottom are open (still with $u = 0$) and no-slip is prescribed on the sides.

In the FS and NS case the null space of the pressure will need to be addressed and we require that the average pressure over the domain is zero, i.e.

$$\iiint_\Omega p(x, y, z) dx dy dx = 0$$

The following quantities are reported:

- the root mean square velocity $\boldsymbol{v}_{rms}$ over the whole domain;

- the minimum and maximum velocity and pressure in the domain (i.e. $u_{min,max}$, $v_{min,max}$, $w_{min,max}$ and $p_{min,max}$;

- the velocity in the center of the sphere (maybe).

The factors which are expected to influence these measurements are:

- the resolution, especially if hexahedral elements are used;

- the quadrature rule, especially if the material properties are directly prescribed on these;

- the type of numerical method and their order (think $Q_1 \times P_0$ vs $Q_2 \times Q_1$ vs $Q_2 \times P_{-1}$ vs ... for finite elements)

- whether full or reduced densities are used (except for OT case);

- the parameter m which controls the rigidity of the sphere with regards to the surrounding fluid;

- the relative density difference between the fluid and the sphere.

Stokes' law was derived by George Gabriel Stokes in 1851. It describes the frictional force a sphere with a density different than the surrounding fluid experiences in a laminar flowing viscous medium around it. By equating the frictional force term $6\pi\eta_f R\nu_s$ with the buoyancy force $4/3\pi R^3 \delta\rho g$, we arrive at the following settling velocity:

$$\nu_s = \frac{2}{9}\frac{\delta\rho R^2 g}{\eta_f}$$

All the measurements above will then be adimensionalised by dividing all velocities by ν_s and pressures by $p_{ref} = \rho_f g L_z = 1$. Note that Stokes law is derived in an infinite fluid so that the recovered sphere velocity measurements are not expected to match this analytical value exactly. In our case we have

$$\nu_s = \frac{2}{9}\frac{0.01 \cdot 0.123456789^2 \cdot 1}{1} \simeq 0.00003387017 = 3.387017 \cdot 10^{-5}$$

Note that as noted in the Gale manual[4] a correction can be made to this velocity when the sphere is itself viscous, see problem 2 p65-66 of second English edition of 'Fluid Mechanics' by Landau & Lifshitz, volume 6 of Theoretical Physics.

---

[4]https://geodynamics.org/cig/software/gale/
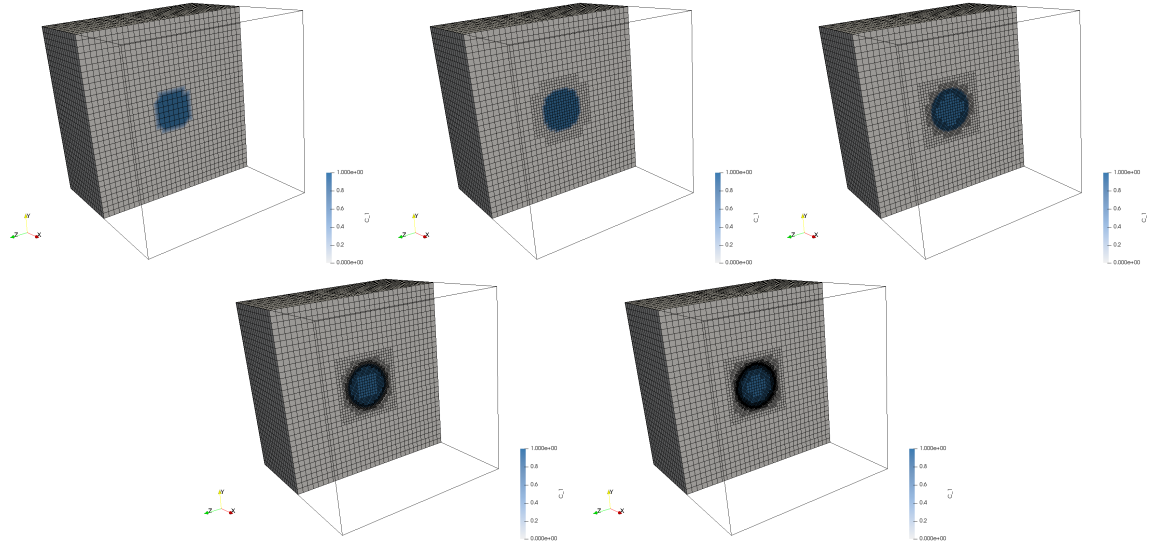
Octree-based ASPECT meshes. The background mesh is 16^3 and refinement is allowed to take place 4 times. Note that a special output is automatically generated in the code which subdivides all elements in 8 for visualisation purposes only.
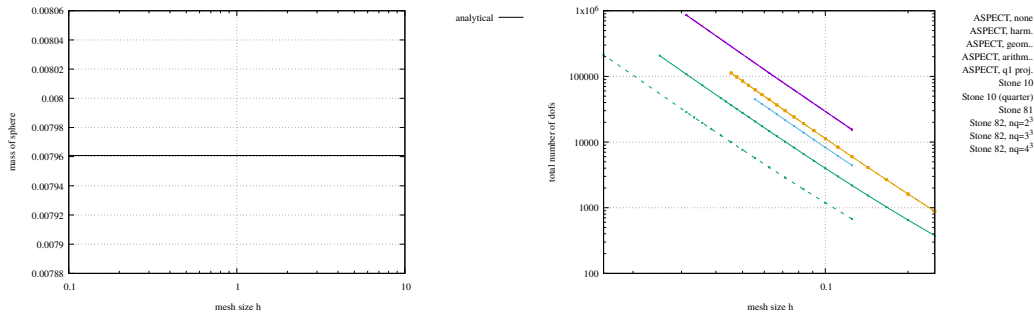
FS results
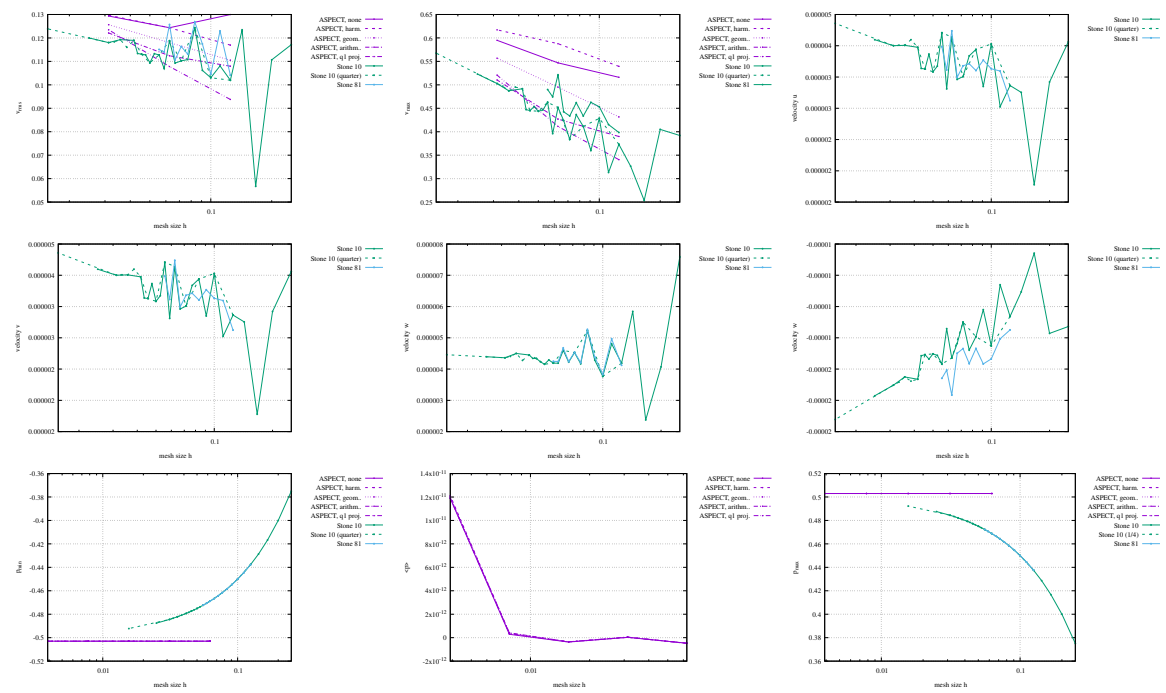
Measurements obtained with ASPECT and STONE 10 for various averaging schemes.
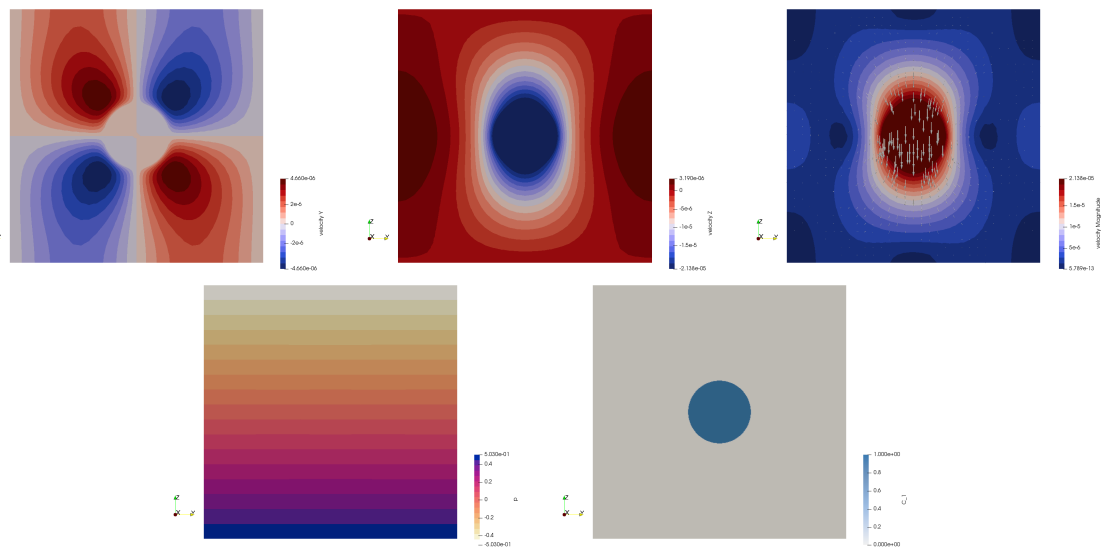
High-resolution solution cross section at $x = 0.5$

NS results
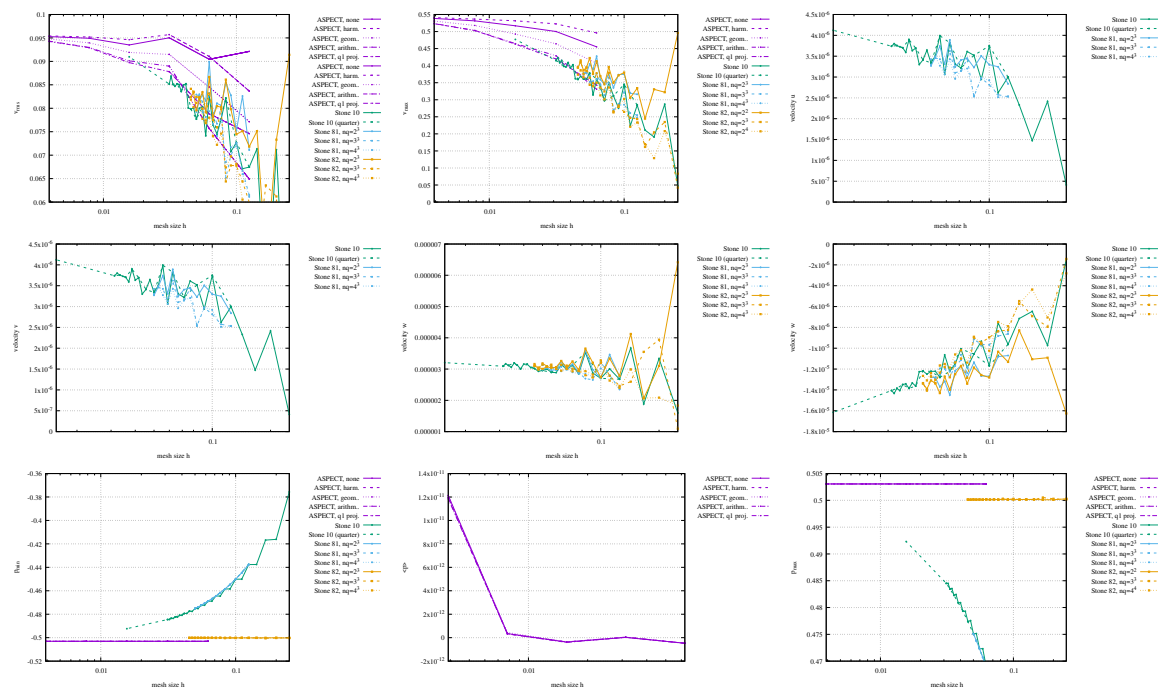
Measurements obtained with ASPECT and STONE 10 for various averaging schemes.

4+4 mesh

1090
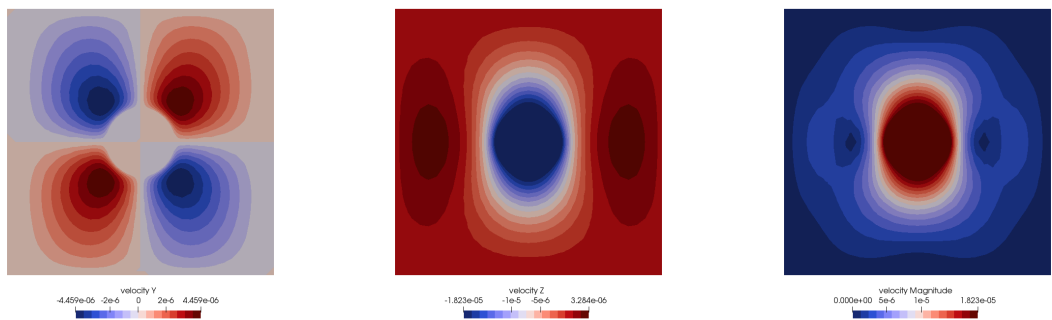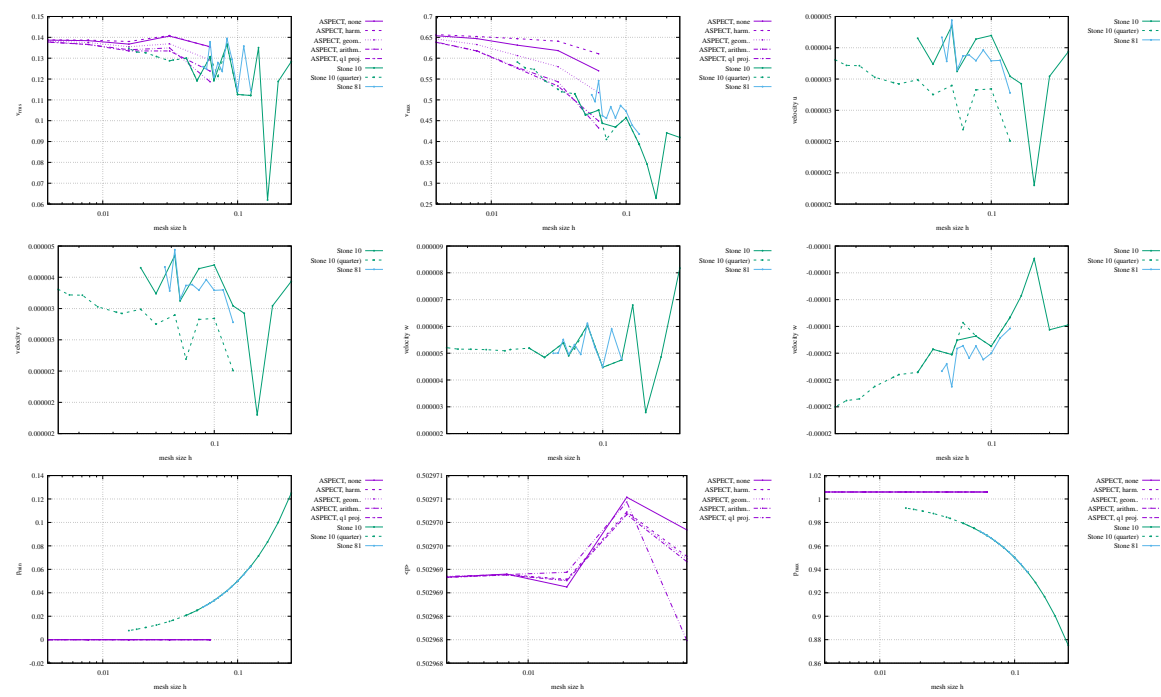
OT results

Measurements obtained with ASPECT and STONE 10 for various averaging schemes.
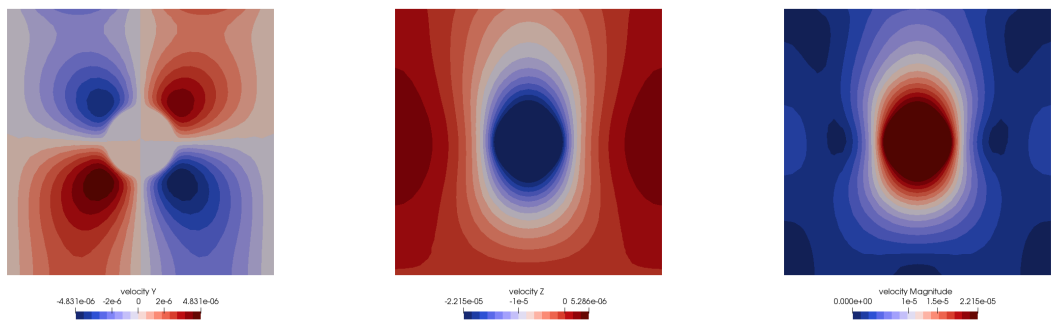
4+4 mesh

CYL results .

note that sphere and walls are only 1000 times more viscous than fluid
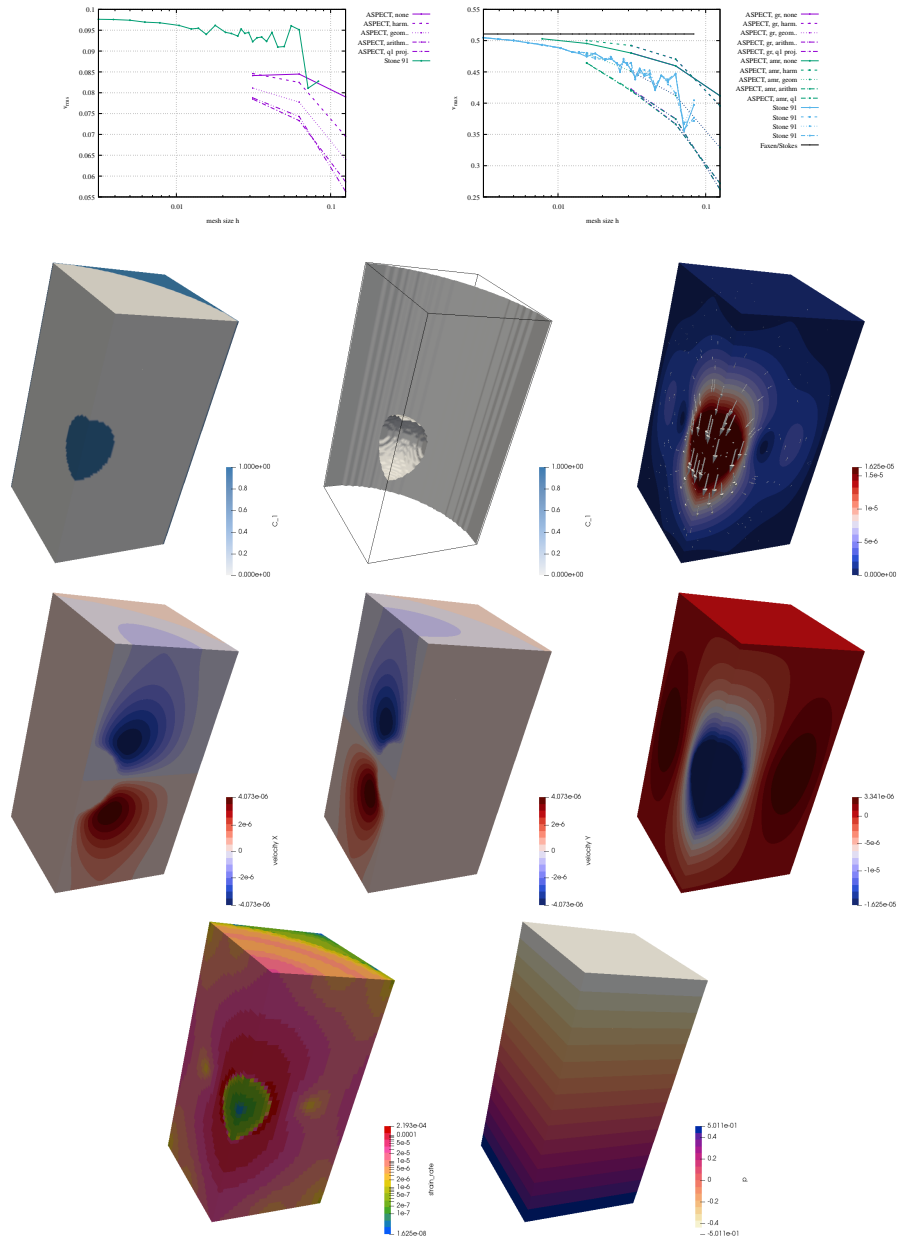
level 5 mesh, 32x32x64 elements.

I have proven in STONE 92 that the sphere should probably be 10^6 times more viscous than the fluid and the box should be 1.5 in height to recover the Habermann/Faxen velocities.
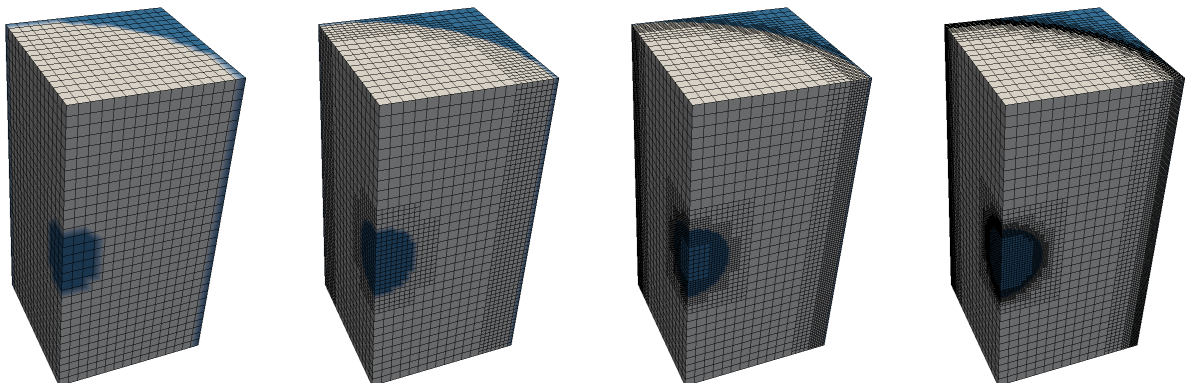
20.24 (Instantaneous) Stokes sphere (2D)

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/stokes_sphere2D

This is the same experiment as in the 3D case but in 2D. When using ASPECT , we simply start with regular meshes ranging from 8^2 to 512^2 elements and we use the default $Q_2 \times Q_1$ element. This corresponds to

```
subsection Mesh refinement
set Initial adaptive refinement = 0
set Initial global refinement = 3->9
set Refinement fraction = 0.9
set Coarsening fraction = 0
set Strategy = composition
end
```
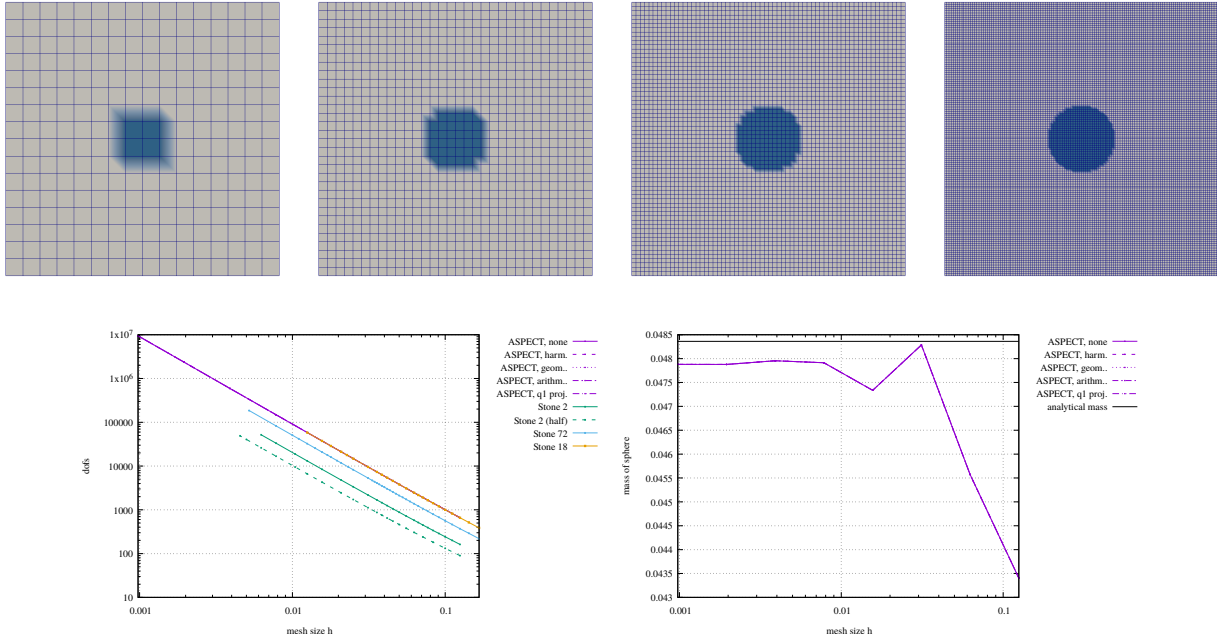

Left: total number of dofs in the Stokes problem; Right: mass of composition 1 as measured in ASPECT .

The total mass of the system is

$$
\begin{align}
M & = (L_x L_y - \pi R^2)\rho_{fluid} + \pi R^2 \rho_{sphere} \tag{20.100} \\
& = (L_x L_y - \pi R^2)\rho_{fluid} + \pi R^2(\rho_{fluid} + \delta\rho) \tag{20.101} \\
& = L_x L_y \rho_{fluid} + \pi R^2 \delta\rho \tag{20.102} \\
& = 1 + \pi \cdot 0.123456789^2 \cdot 0.01 \tag{20.103} \\
& \simeq 1.00047882831 \tag{20.104}
\end{align}
$$

The Stokes velocity can be obtained as follows: on p61 of Landau & Lifschitz, it is reported that the drag force on a disk moving in its plane is $F = 32\eta_f R \mathbf{v}_s/3$. The buoyancy force is $F = \pi R^2 \delta\rho g$, so the velocity is then

$$\mathbf{v}_s = \frac{3\pi}{32}\frac{\delta\rho}{\eta_f}Rg \simeq 0.00036361025$$

Given the dimensions, this is obviously given per meter of the infinite cylinder. This is substantially smaller than what we recover, so I keep the 3D velocity as reference for now.

In a second time, we make use of the mesh refinement capabilities of the code, as shown here:
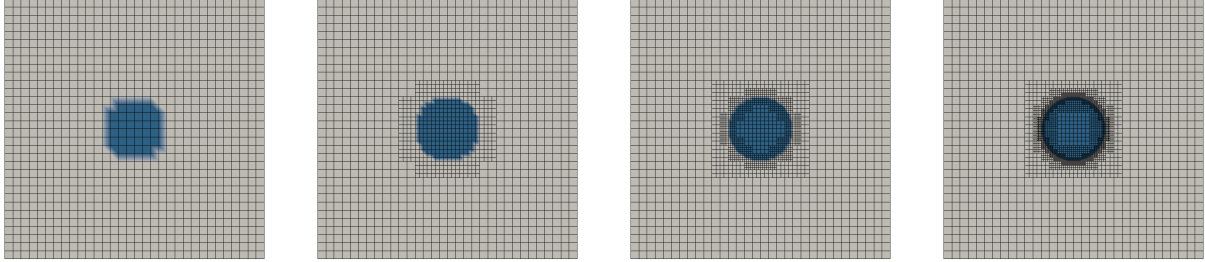
Octree-based ASPECT meshes. The background mesh is 16^3 and refinement is allowed to take place 4 times. Note that a special output is automatically generated in the code which subdivides all elements in 8 for visualisation purposes only. Initial refinements 0,1,2,3,4.

This corresponds to

```
subsection Mesh refinement
set Initial adaptive refinement = 0 -> 9
set Initial global refinement = 4
set Refinement fraction = 0.9
set Coarsening fraction = 0
set Strategy = composition
end
```

FS results
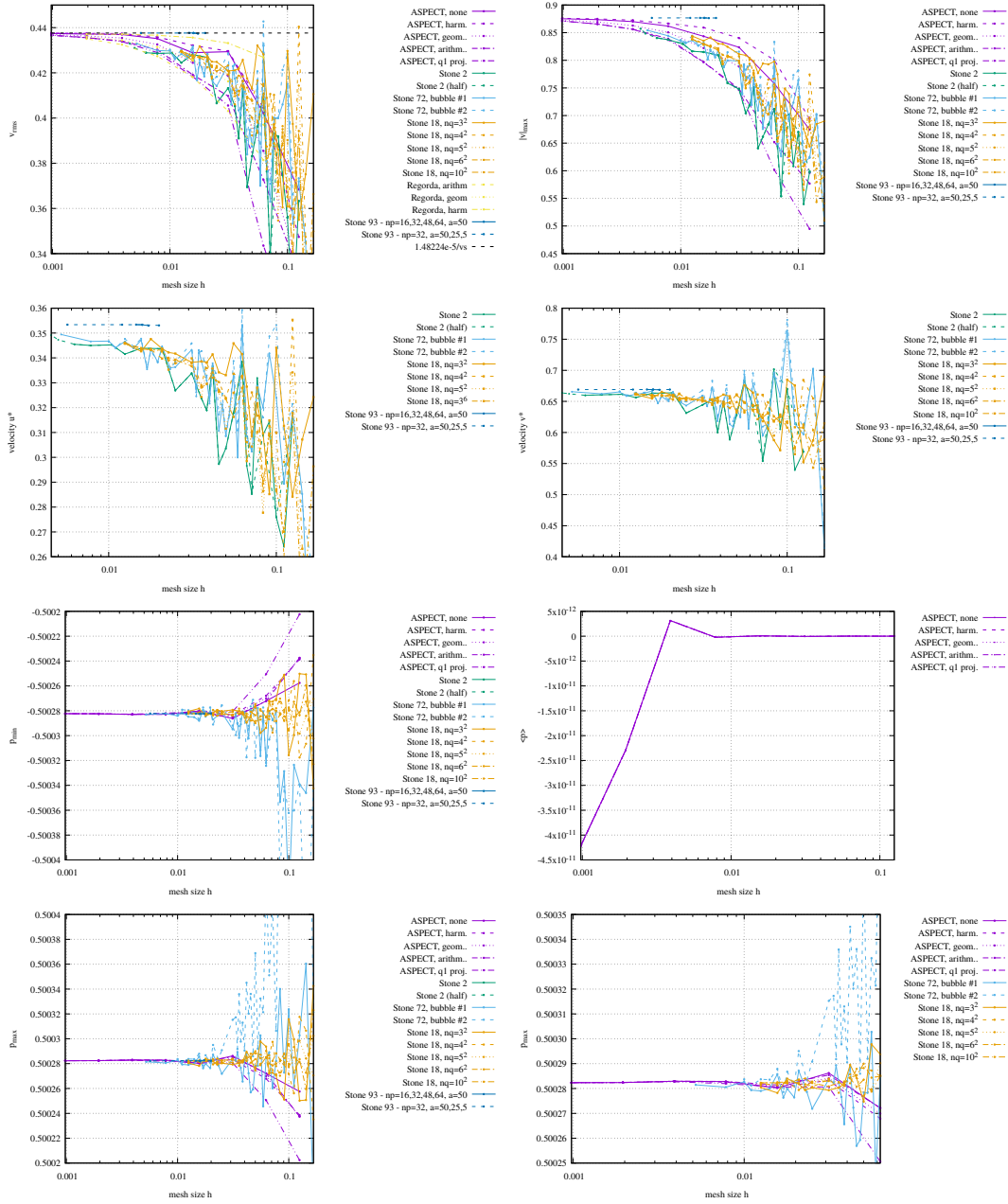
Measurements obtained with ASPECT for various averaging schemes and with different stones.

1095
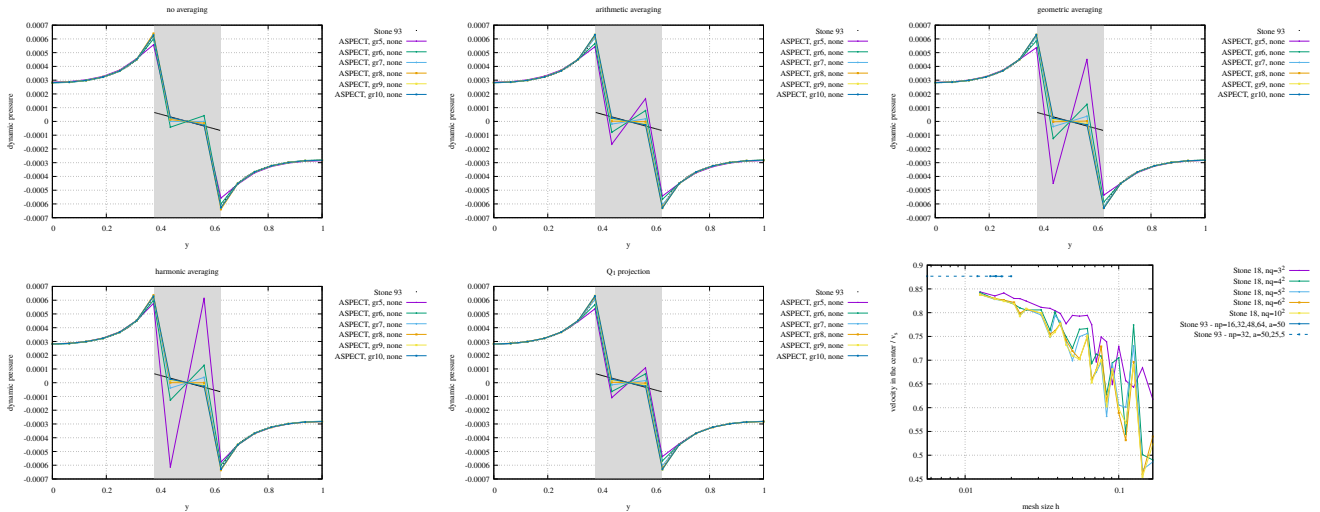
NS results
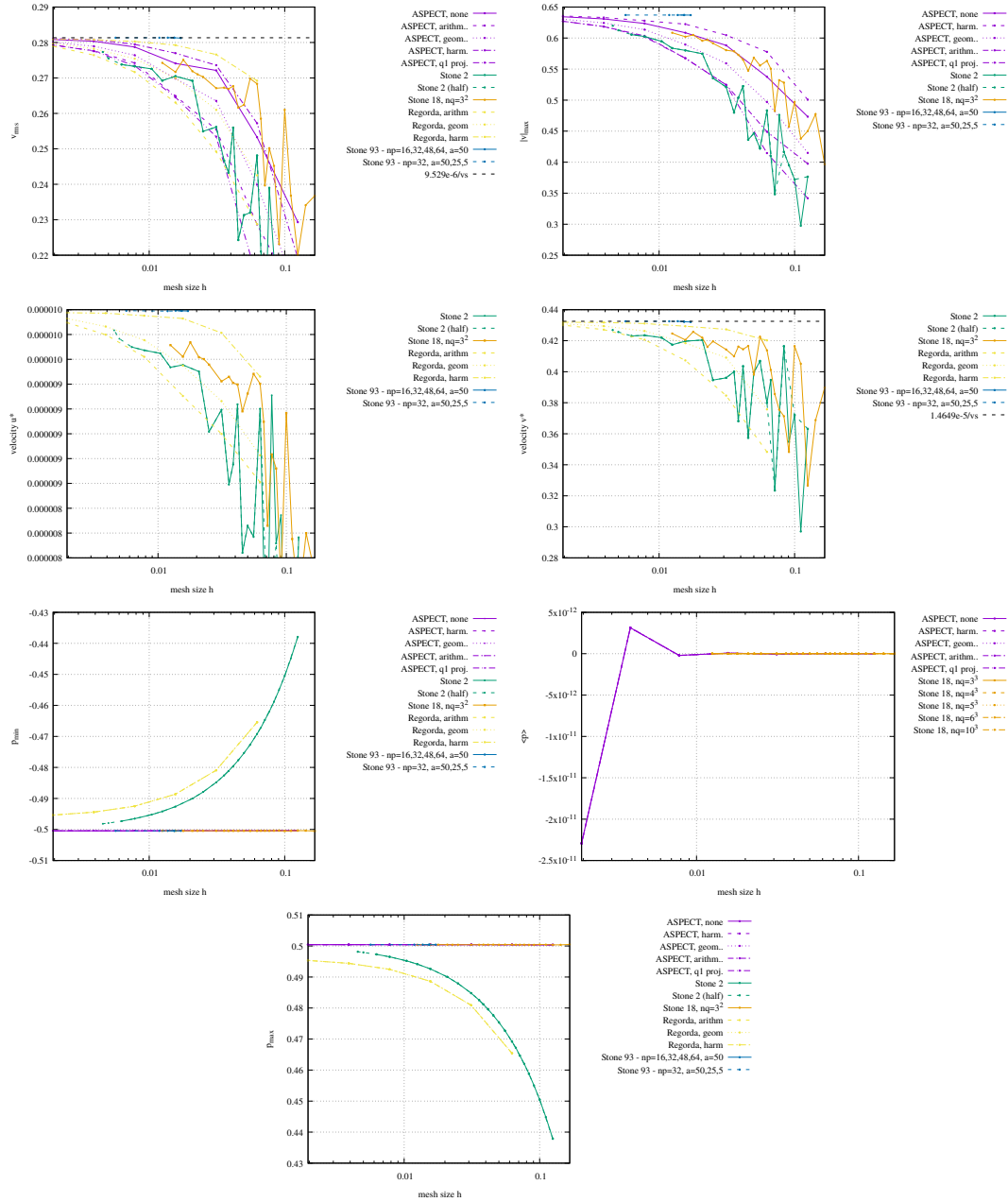
Measurements obtained with ASPECT for various averaging schemes and with different stones.

I have also retrieved the pressure at 16 equidistant locations on the $x = 0.5$ line for all five averagings. Because the signal is dominated by the lithostatic pressure I have subtracted it from the data, so that I hereunder plot the dynamic pressure as a function of the y coordinate:
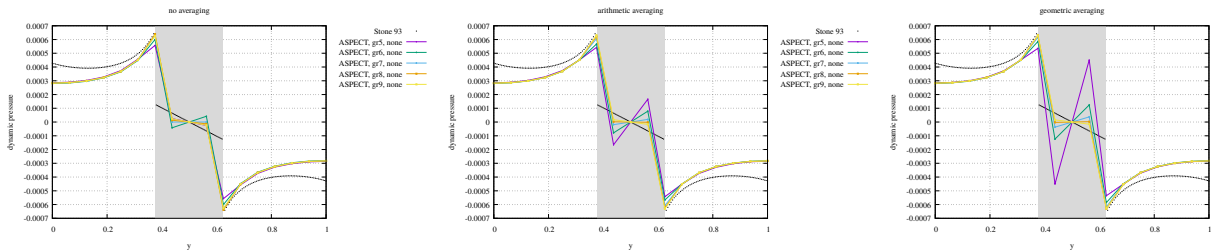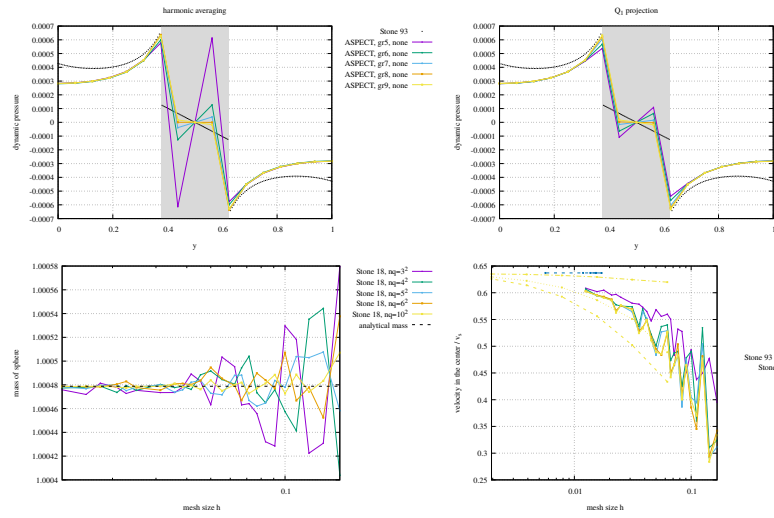
OT results
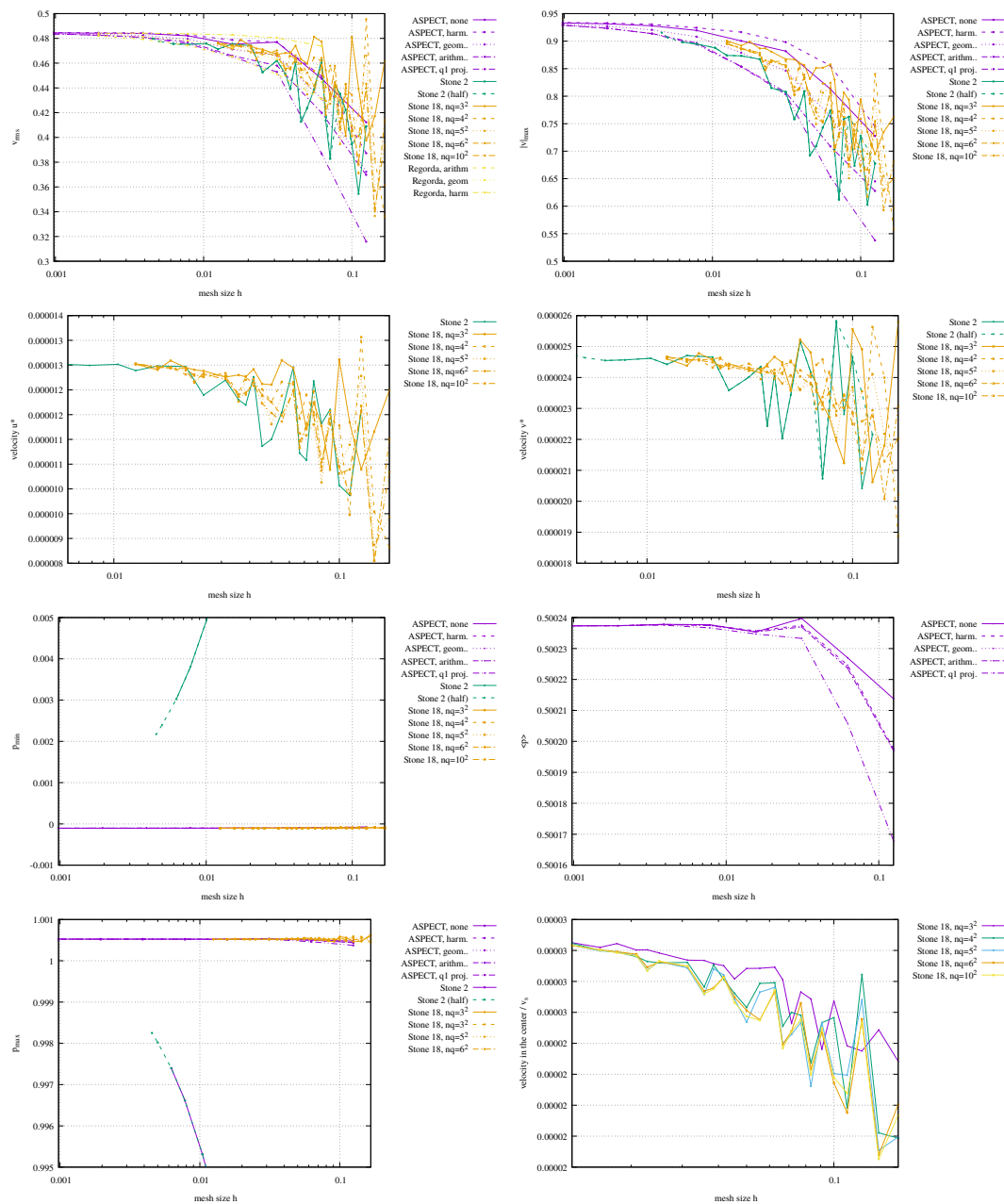
Measurements obtained with Aspect for various averaging schemes and with different stones.
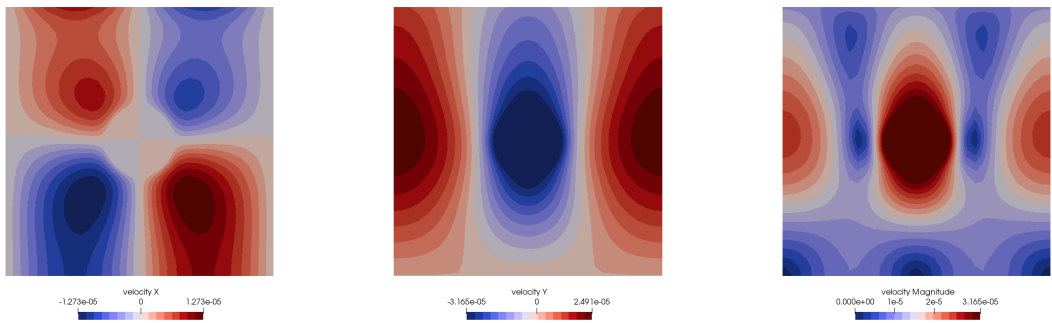
BO results .
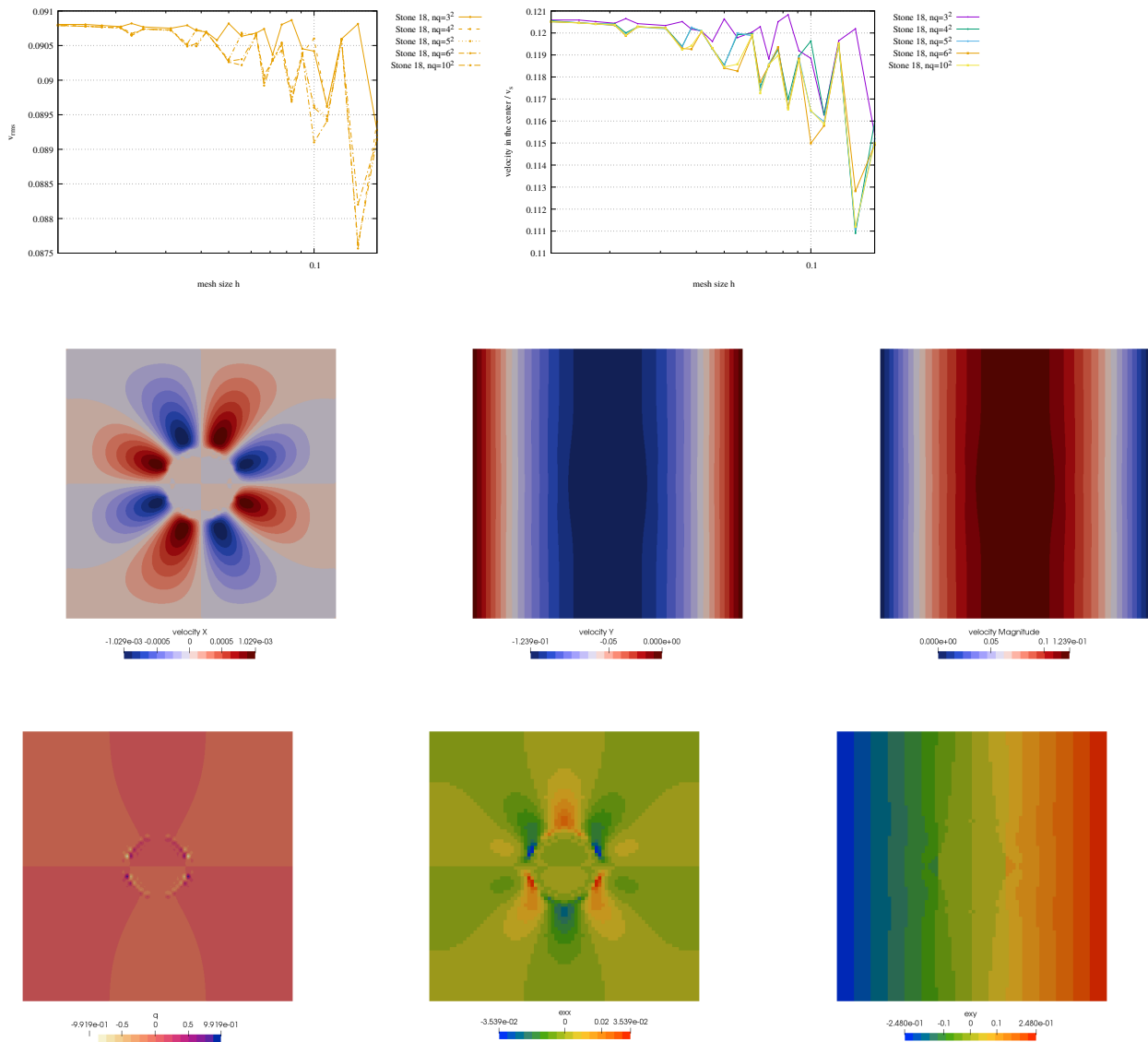
20.25 Stokes sphere (2D) in fluid with deformable free surface

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/stokes_sphere_fs2D/

The domain is a 1×0.75 box. If sticky air is used, then its thickness should be 0.25 so that the domain is a unit square. The fluid is characterised by $\rho_f = 1$ and $\eta_f = 1$. The sphere is characterised by $\rho_s = 2$ and $\eta_s = 10^3$. The air is characterised by $\rho_a = 0$ and $\eta_a = 10^{-3}$. Gravity is vertical with $\vec{g} = -\vec{e}_y$. The sphere has a radius $R_S = 0.123456789$ and its center is at position $\vec{r}_c = (0.5, 0.6)$. Boundary conditions are free slip on all sides (unless a true free surface is used). Pressure is normalised so that its average is zero on the top (if no free surface is used). The model is run for 200s. The CFL number is set to 0.25 with a maximum time step of 0.5.

We wish to keep track of the following quantities as a function of time:

- the position and velocity of the sphere center,

- the minimum and maximum topography,

- the volume of fluid $V_f(t)$, sphere $V_s(t)$ and air $V_a(t)$

- root mean square velocity ν_{vrms} for the whole domain, as well as for the air, fluid and sphere separately, and for the fluid+sphere,

- the maximum velocity and pressure in the domain,

- the time step value δt,

- the average density[5] and viscosity in the domain:

$$
\begin{aligned}
\langle \rho \rangle(t) &= \frac{1}{L_x L_y} \iint \rho(x, y, t) dx dy = \frac{1}{L_x L_y}(V_a(t)\rho_a + V_f(t)\rho_f + V_s(t)\rho_s) \\
\langle \eta \rangle(t) &= \frac{1}{L_x L_y} \iint \eta(x, y, t) dx dy \frac{1}{L_x L_y}(V_a(t)\eta_a + V_f(t)\eta_f + V_s(t)\eta_s) \quad (20.105)
\end{aligned}
$$

Initial values are

$$
\langle \rho \rangle(0) = \frac{1}{L_x L_y}(V_a(0)\rho_a + V_f(0)\rho_f + V_s(0)\rho_s) = 0.25*0 + (0.75 - \pi R_s^2)*1 + \pi R_s^2*2 = 0.75 + \pi R_s^2 \simeq 0.79788283
$$

$$
\langle \eta \rangle(0) = \frac{1}{L_x L_y}(V_a(0)\eta_a + V_f(0)\eta_f + V_s(0)\eta_s) = 0.25*10^{-3} + (0.75 - \pi R_s^2)*1 + \pi R_s^2*10^3 \simeq 48.5851989989
$$

- the min/max of the compositional fields when these are used;

- the velocity, pressure and material at position (0.5,0.6);

---

[5]Because $L_x L_y = 1$, also equal to the total mass of the system

- the pressure at position (0.5,0).

Participating codes:

- ASPECT uses $Q_2 \times Q_1$ elements by default. $Q_2 \times P_{-1}$ elements can be used by setting

```
subsection Discretization
  set Use locally conservative discretization = true
end
```

The default stabilisation method when compositional fields are used is the entropy viscosity method, but SUPG has also been implemented and can be triggered with

```
subsection Discretization
  subsection Stabilization parameters
    set Stabilization method = SUPG
  end
end
```

The default mesh settings are as follows:

```
subsection Mesh refinement
  set Initial adaptive refinement  = 1
  set Initial global refinement    = 6
  set Refinement fraction          = 0.9
  set Strategy                     = composition
  set Coarsening fraction          = 0.1
end
```

In the results hereafter when the combination 6-0 or 7-0 are mentioned, this means that the coarsening fraction has been set to zero and these correspond then to regular meshes with 64x64 and 128x128 elements respectively.

Active particles can also replace the compositional fields and this is how it is triggered from the input file:

```
subsection Compositional fields
  set Number of fields = 2
  set Names of fields = sphere, air
  set Compositional field methods = particles, particles
  set Mapped particle properties = sphere:initial sphere, air: initial air
end

subsection Postprocess
  set List of postprocessors = visualization, ... , particles
  subsection Visualization
    set List of output variables = density, viscosity, strain rate
    set Time between graphical output = 1
  end
  subsection Particles
    set Number of particles = 100000
    set Time between data output = 0
    set Data output format = vtu
    set List of particle properties = velocity, initial composition, initial
  position #, integrated strain
    set Interpolation scheme = cell average
    set Update ghost particles = true
    set Particle generator name = random uniform
  end
end
```

- Stone 93. Code based on unstructured mesh of Crouzeix-Raviart triangular elements. The resolution is controlled by the minimum area of the triangles as passed as argument to the triangle mesher, and the parameter np which controls the number of points on the hull (np on each side), the surface ($5 * np$) and the sphere ($5 * np$);

- Alessandro Regorda's code: The number of markers is fixed per element with random distribution. At the beginning of the simulation there are: 562500 for 150x150 grid with 25 markers per element, 600,000 for 200x200 grid with 15 markers per element, 655,360 for 256x256 grid with 10 markers per element, 2,621,440 for 512x512 grid with 10 markers per element.

  The advection is RK4. The code maintains the number of markers per element between half of the initial number and the initial number plus half (e.g. in 512x512 markers are between 5 and 15). When in an element there are less markers than the minimum it adds random markers to reach the minimum, while if the number is higher than the maximum some of them are deleted. In this way elements are never empty. When new markers are added they assume the type of the nearest marker. Averaging only applies to viscosity, density is always arithmetically averaged.

In what follows 6-0 and 7-0 correspond to regular grids (no coarsening, no refinement)

Root mean square velocity
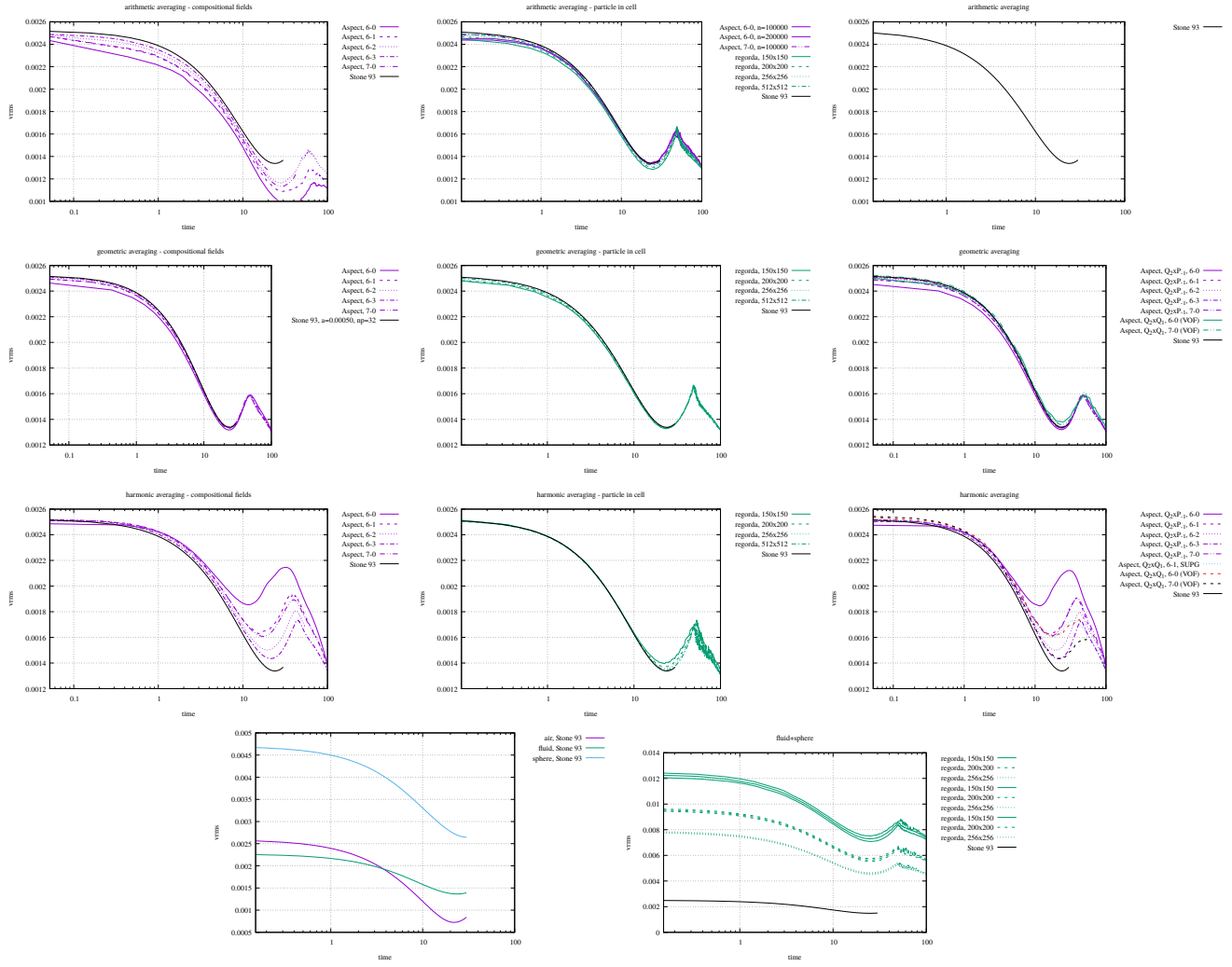
Volumes
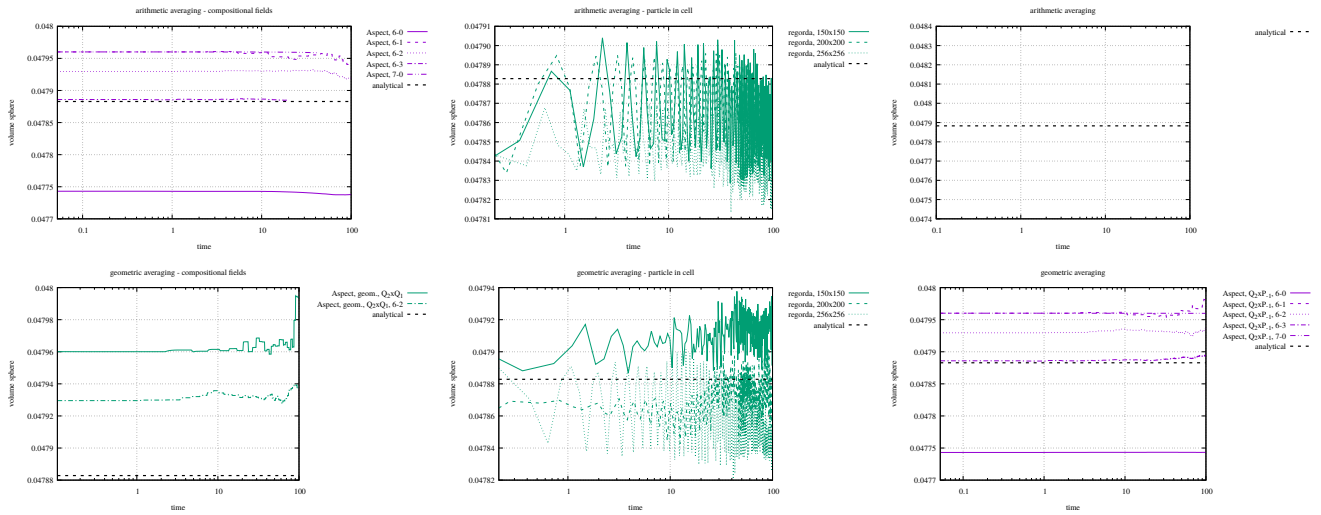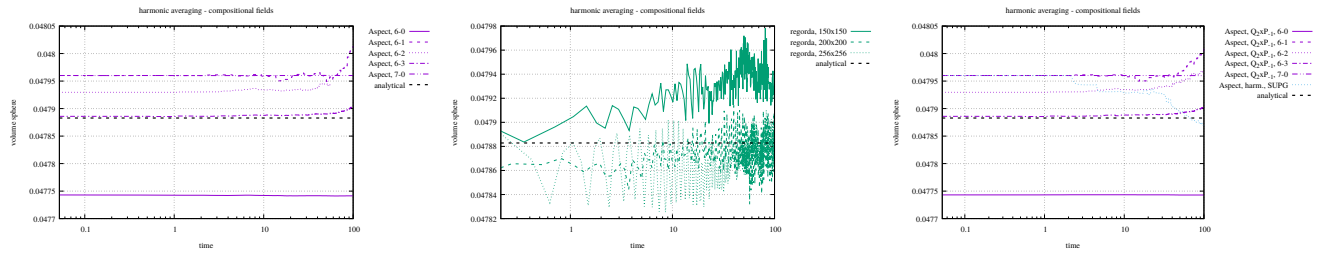
Maximum velocity
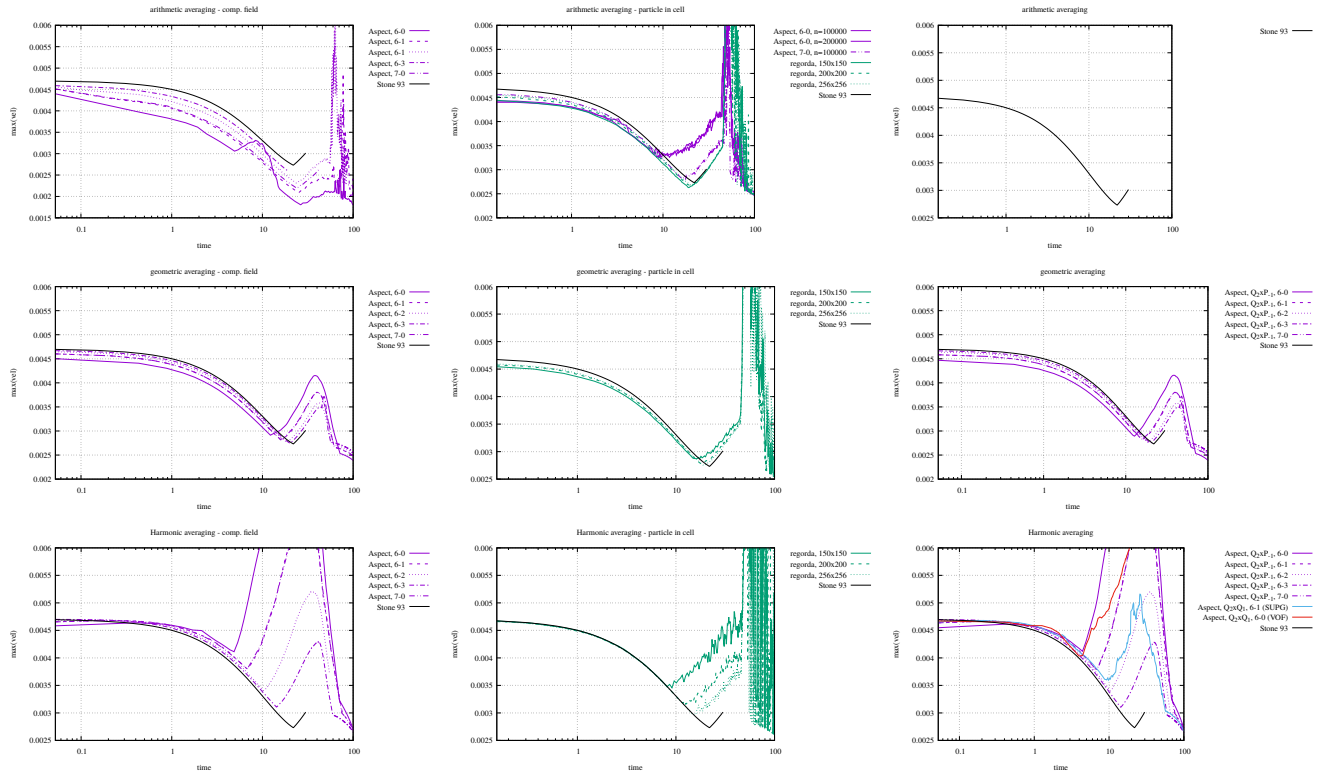
Compositions min/max

1104
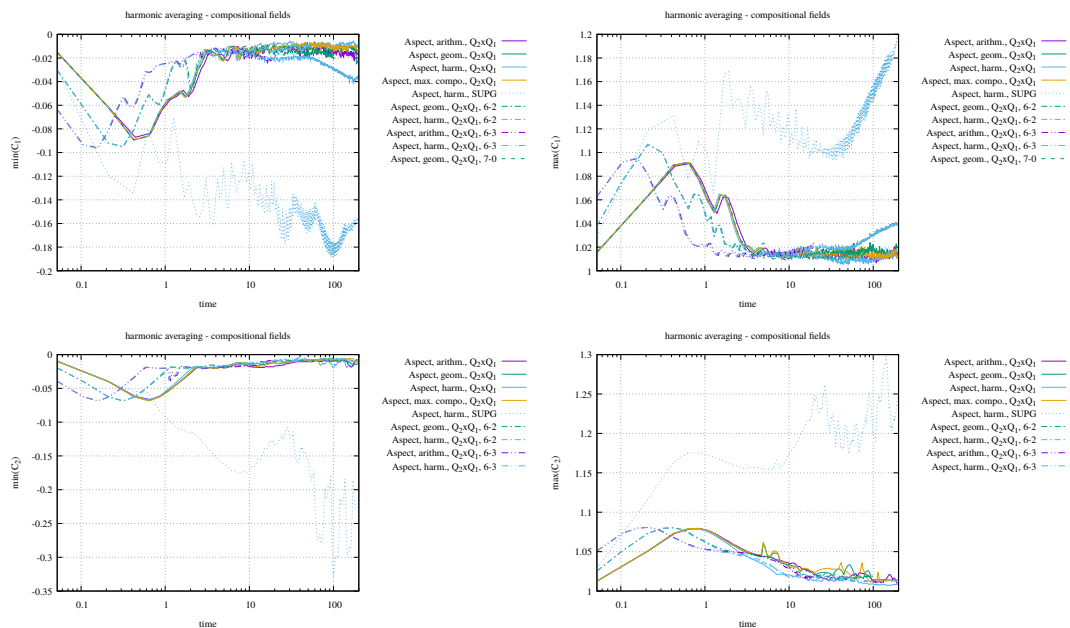
Viscosity volume average
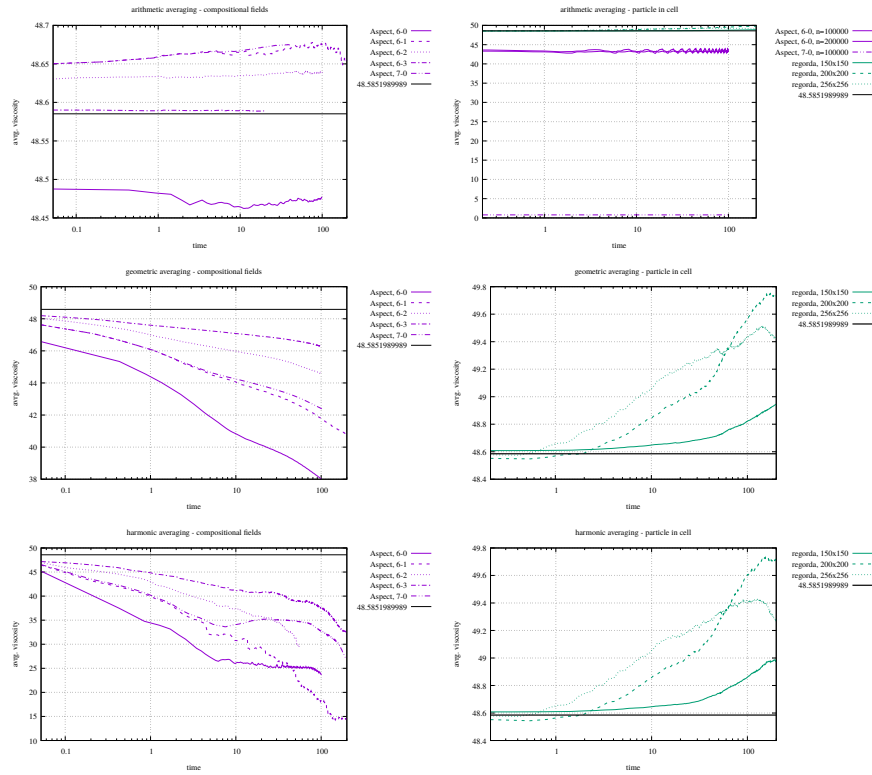
Density volume average
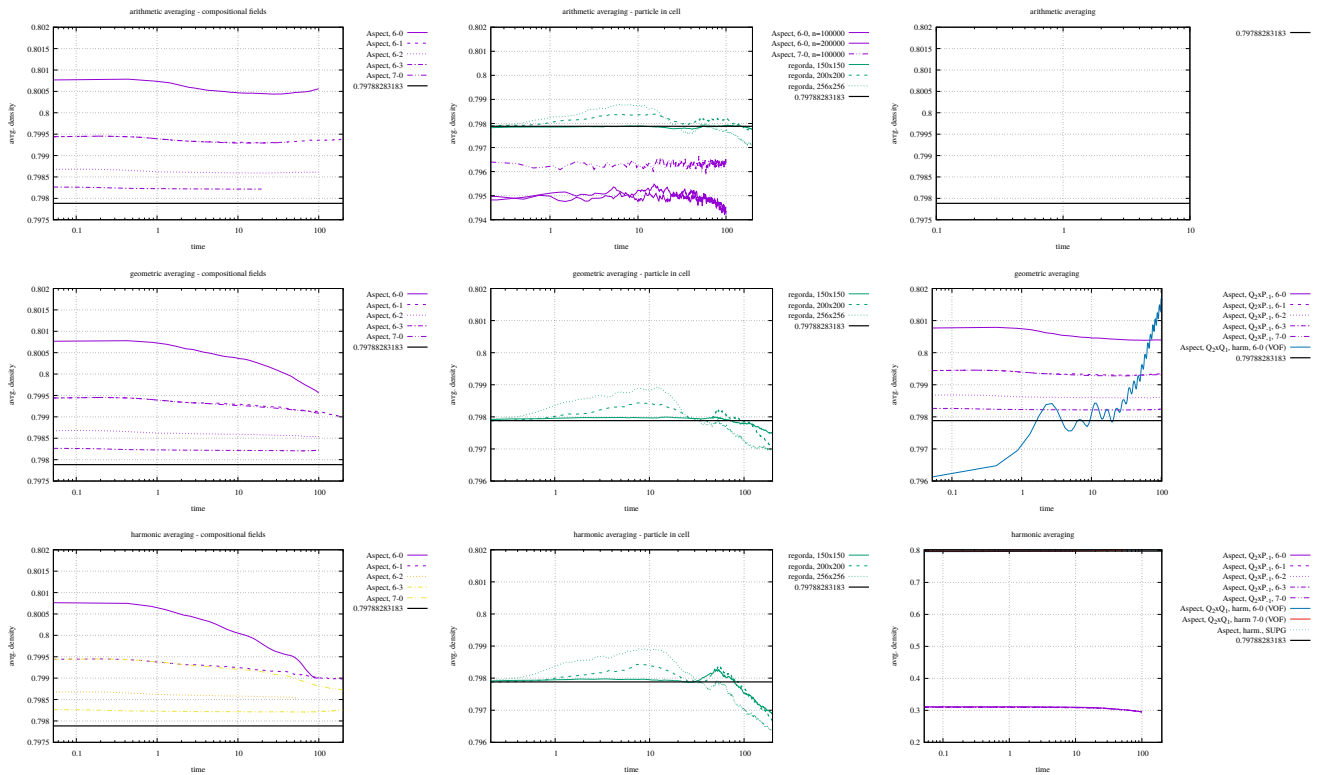
Topography
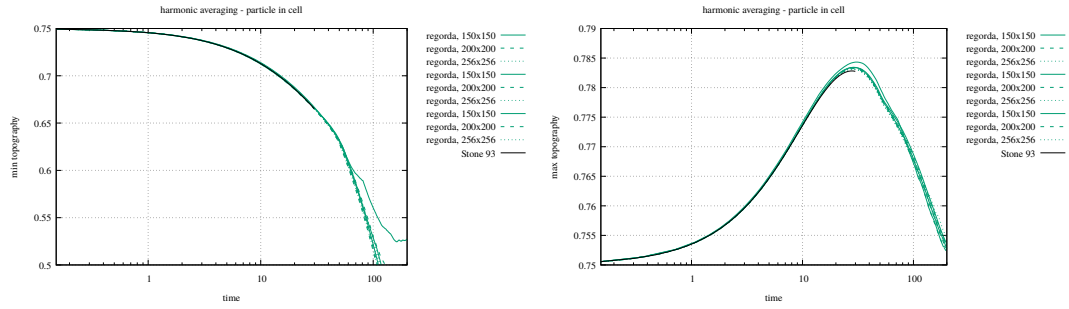
Pressure measurements
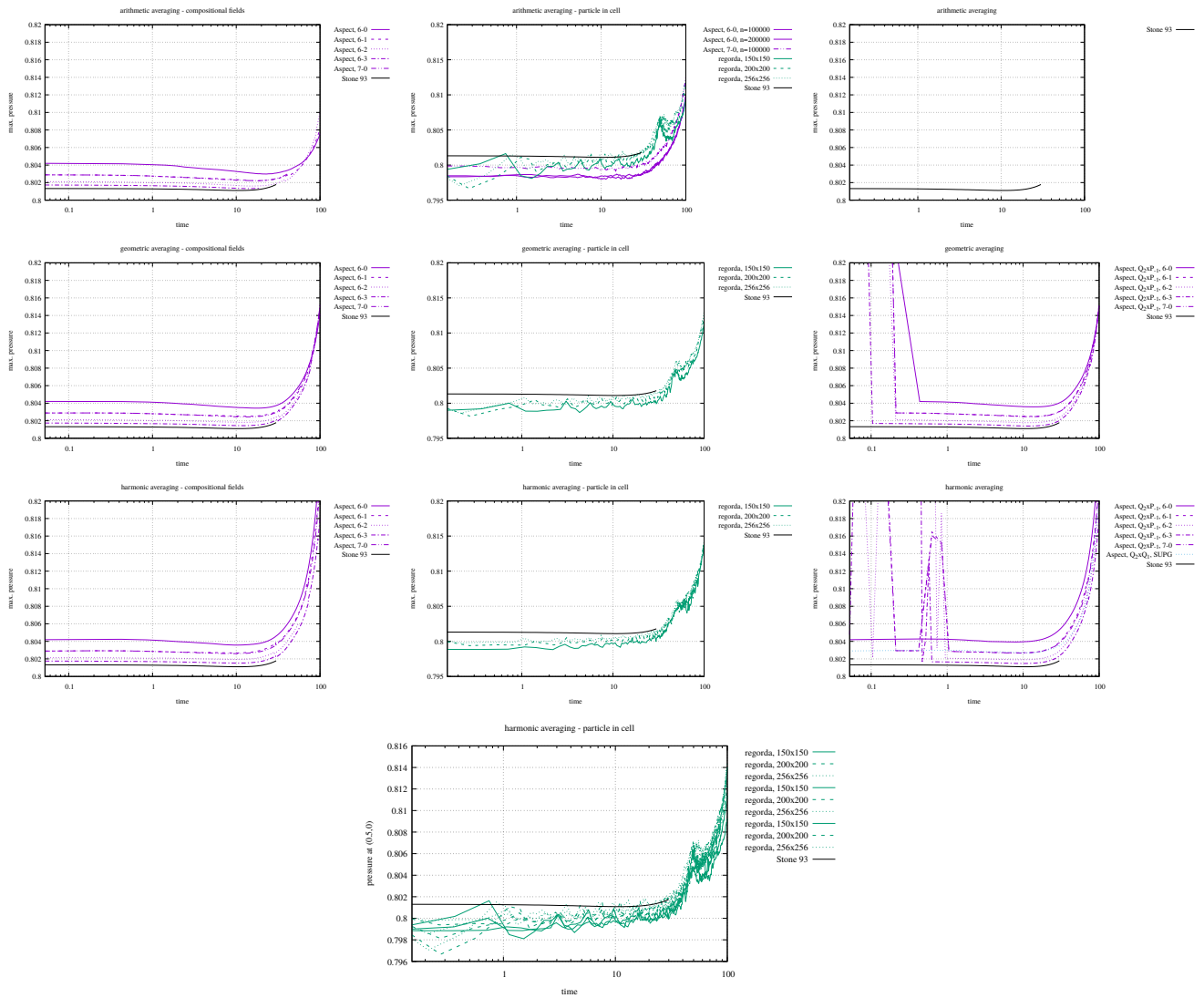
Position of the sphere center .
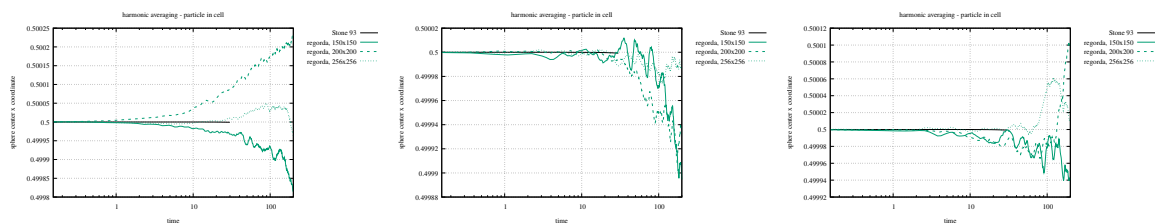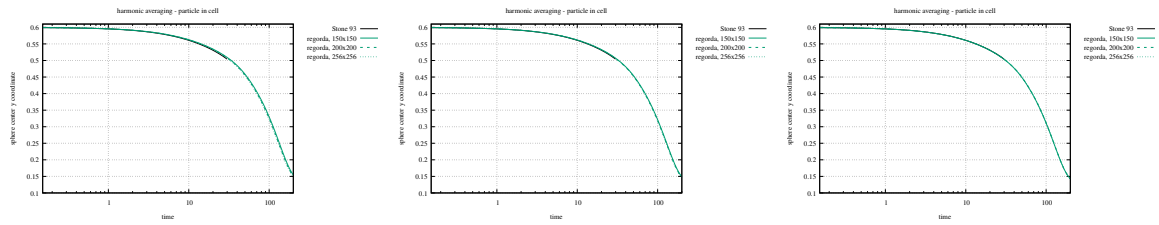
Velocity of the sphere center .
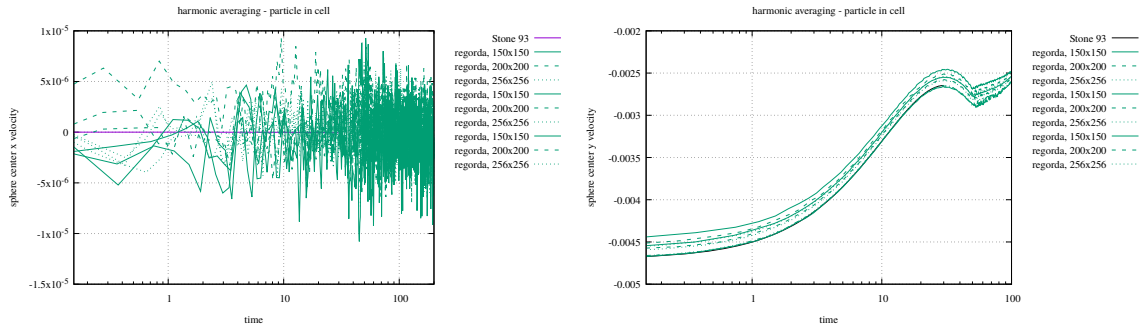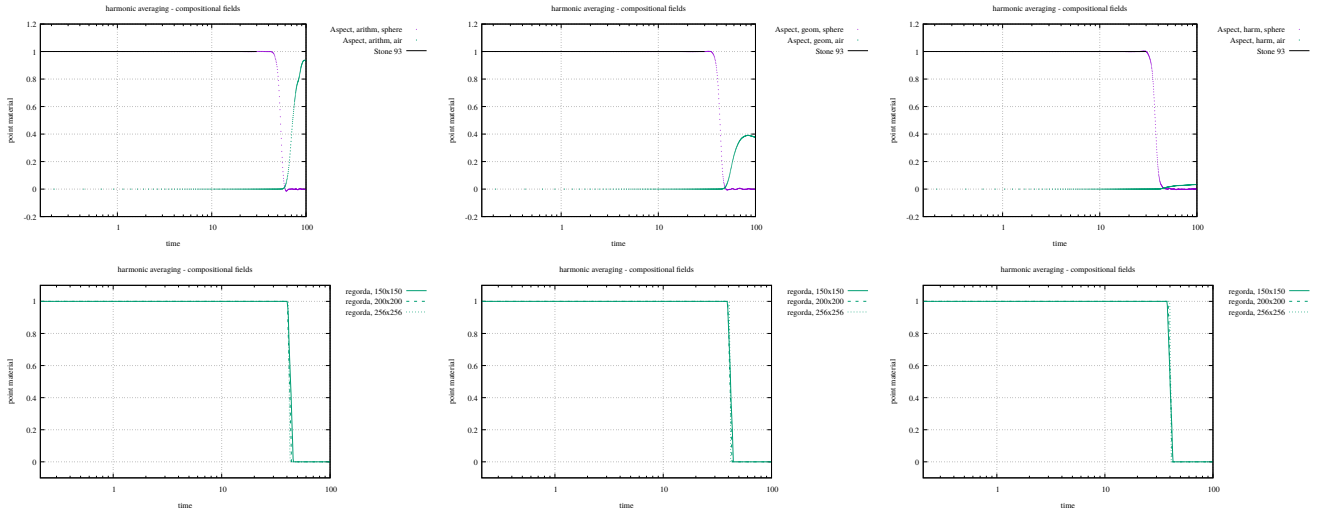

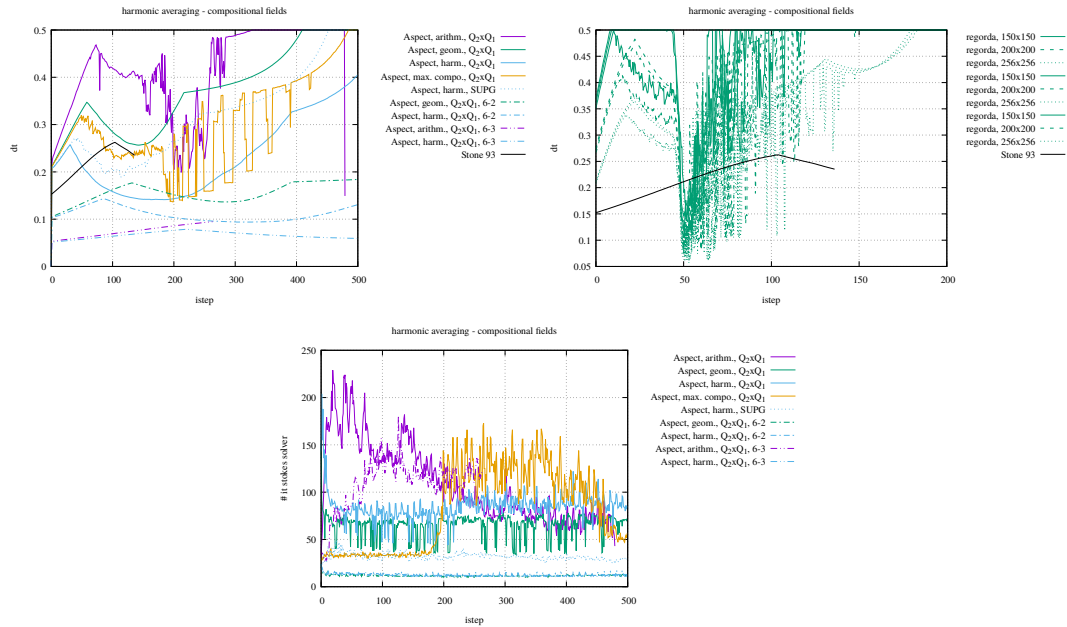**Velocity at (0.5,0.6) location** .
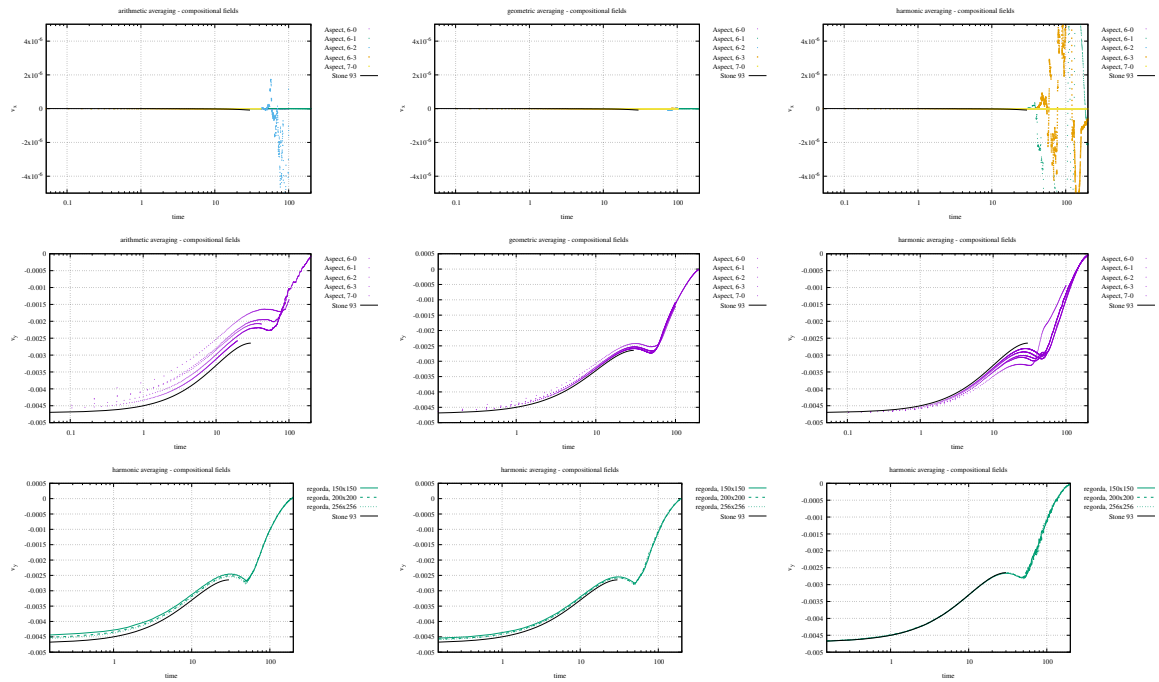
Pressure at (0.5,0.6) location .
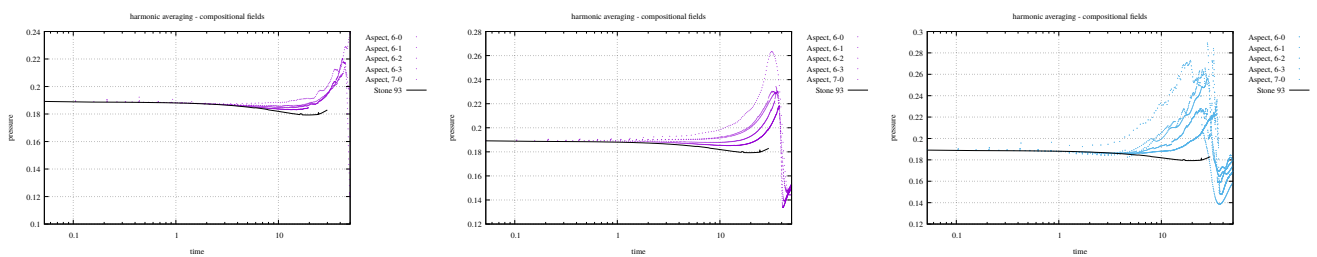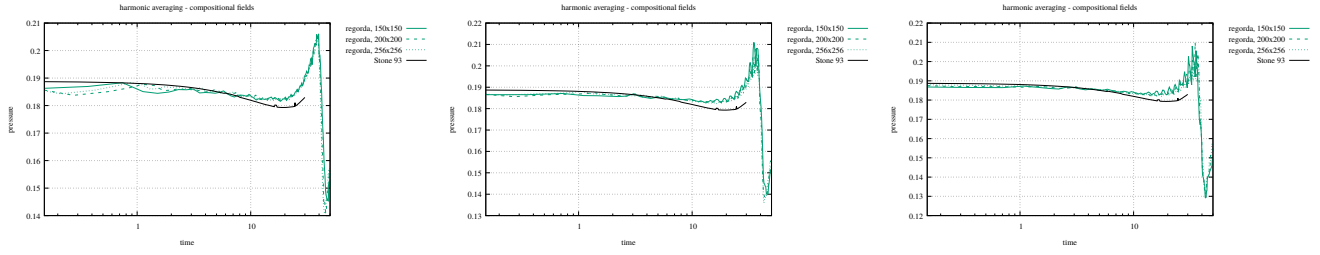
## Material at (0.5,0.6) location .

Timestep and solver convergence

Stone 93 results seem to be most influenced by the resolution on the sphere and surface than the resolution in the fluid. Some conclusions: arithmetic yields very high inner iteration counts. Discontinuous pressure also better on that topic. Geometric averaging yields very good agreement for vrms. Funny enough, geometric does not correspond to a physical arrangement of viscous dampers... Arithm and harm do ultimately converge towards geom but at very high resolution. Using no amr does not change things that much.
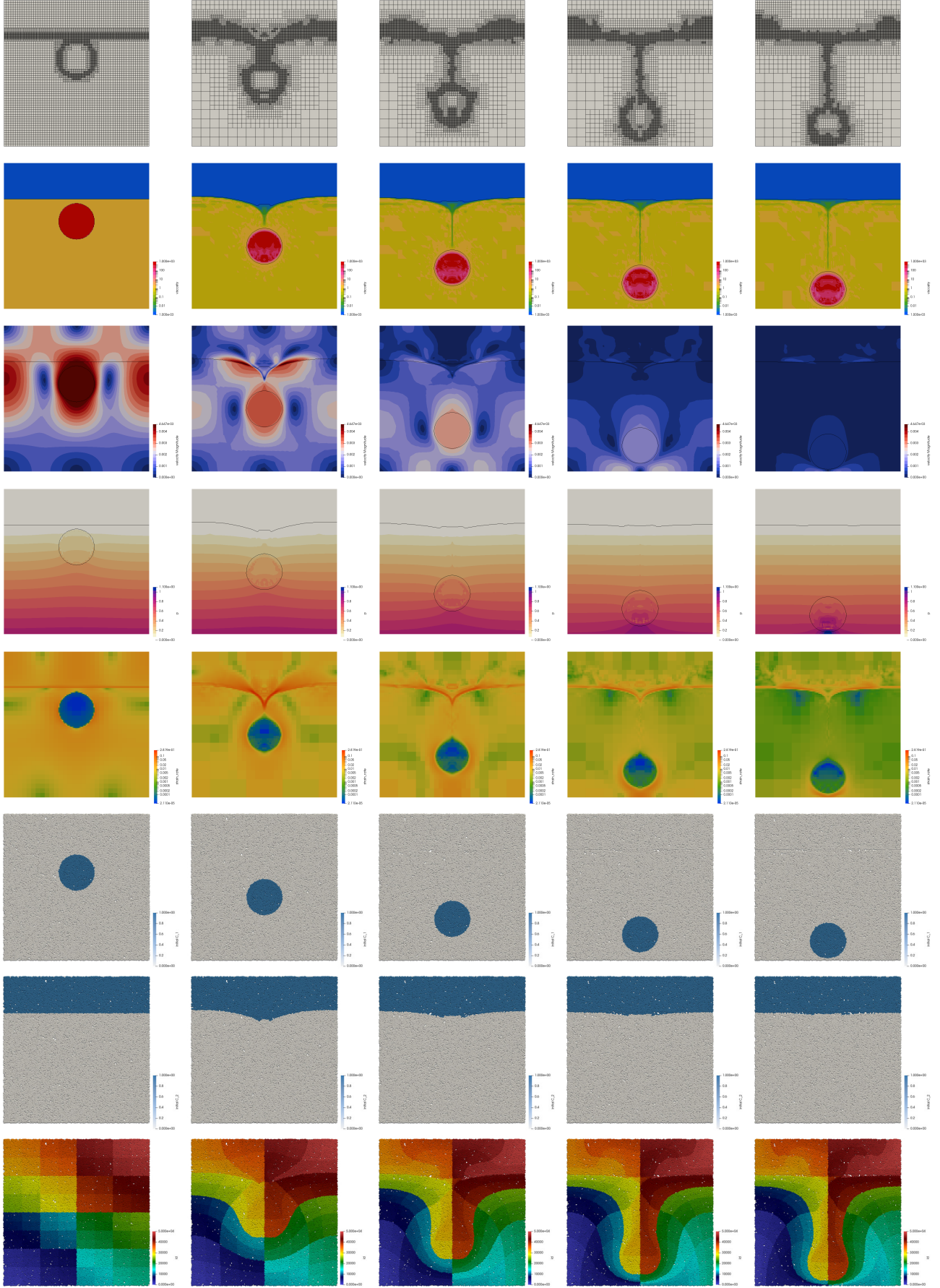
ASPECT results: fields and passive markers time evolution. Last row is the particle id, between 0 and 50000. System at times 0,50,100,150,200. Harmonic averaging.
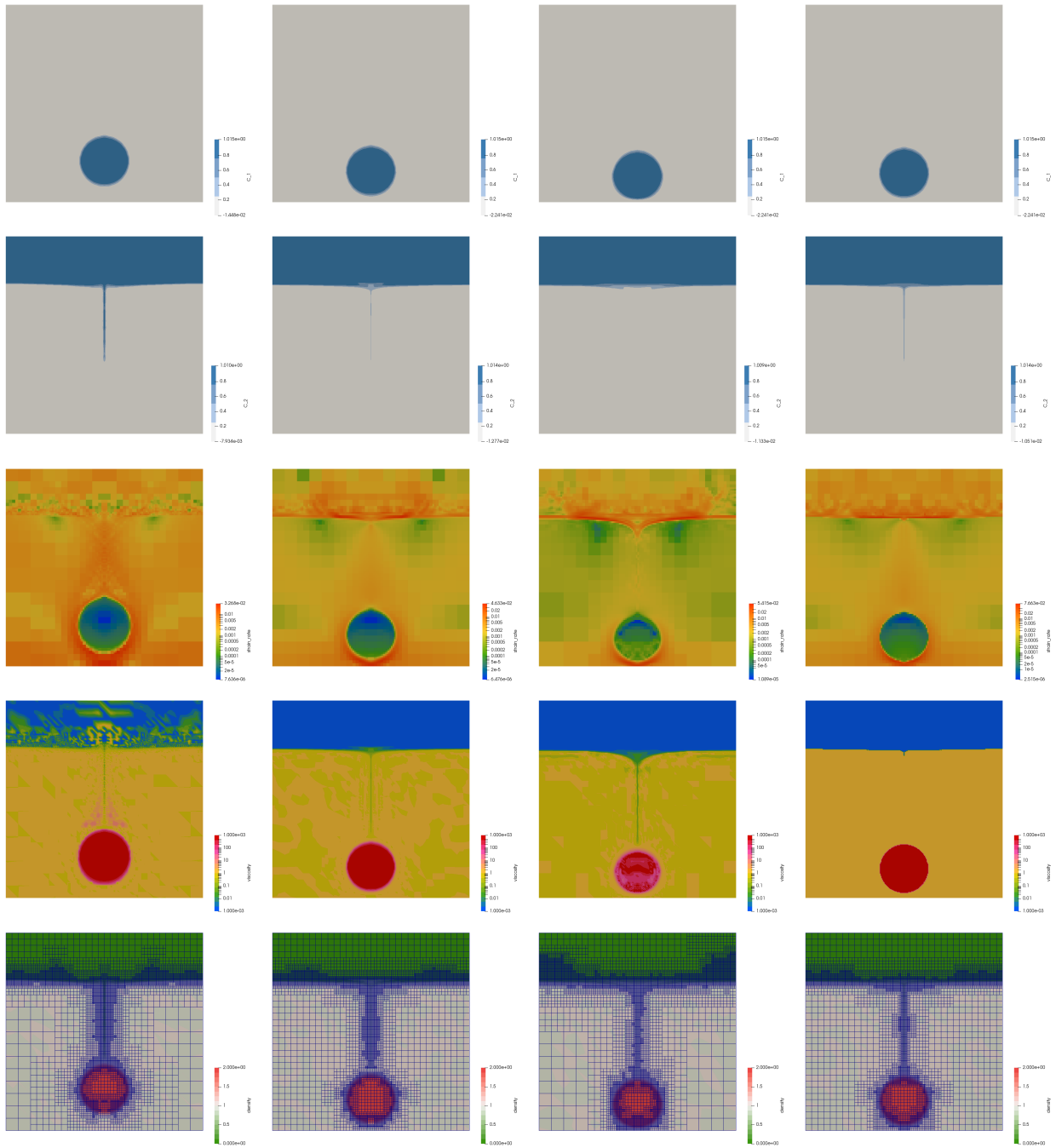
Obtained with Aspect. From left to right: Arithmetic, geometric, harmonic, maximum composition, all at $t = 200$.

This is the ASPECT input file for this benchmark:

```
set Dimension                            = 2
set Start time                           = 0
set End time                             = 200
set Use years in output instead of seconds = false
set CFL number                           = 0.25
set Output directory                     = output-stokes
set Maximum time step                    = 0.5
set Pressure normalization               = surface

subsection Solver parameters
  subsection Stokes solver parameters
    set Number of cheap Stokes solver steps = 0
  end
end

subsection Geometry model
  set Model name = box
  subsection Box
    set X extent  = 1
    set Y extent  = 1
  end
end

subsection Boundary velocity model
  set Tangential velocity boundary indicators = left, right, bottom, top
end

subsection Material model
  set Model name = multicomponent
  subsection Multicomponent
    set Densities = 1, 2, 0
    set Viscosities = 1, 1000, 0.001
    #set Viscosity averaging scheme = maximum composition
    #set Viscosity averaging scheme = arithmetic
    #set Viscosity averaging scheme = geometric
    set Viscosity averaging scheme = harmonic
    set Thermal expansivities = 0
  end
end

subsection Gravity model
  set Model name = vertical
  subsection Vertical
    set Magnitude = 1
  end
end

subsection Boundary temperature model
  set Fixed temperature boundary indicators = bottom, top
  set List of model names = box
end

subsection Initial temperature model
  set Model name = function

  subsection Function
    set Function expression = 0
  end
end

subsection Compositional fields
  set Number of fields = 2
end

subsection Initial composition model
  set Model name = function

  subsection Function
    set Variable names      = x,y
    set Function constants  = r=0.123456789, xc=0.5, yc=0.6
    set Function expression = if(sqrt((x-xc)*(x-xc)+(y-yc)*(y-yc)) < r, 1, 0) ; if (y>0.75,1,0)
  end
end

subsection Mesh refinement
  set Initial adaptive refinement   = 1
  set Initial global refinement     = 6
  set Refinement fraction           = 0.9
  set Strategy                      = composition
  set Coarsening fraction           = 0.1
end

subsection Postprocess
  set List of postprocessors = visualization, velocity statistics, composition statistics, pressure statistics,
    material statistics, global statistics, point values, particles

  subsection Point values
    set Evaluation points = 0.5,0.6
  end

  subsection Particles
    set Number of particles        = 50000
    set Time between data output = 1
    set Data output format        = vtu
    set List of particle properties = initial composition, initial position
  end

  subsection Visualization
    set List of output variables = density, viscosity, strain rate
    set Time between graphical output = 1
  end
```

20.26 Relaxation of topography (Crameri *et al.* , 2012)

This benchmark was first presented in Crameri *et al.* (2012) [304] and is also presented in Hillebrand *et al.* (2014) [604]. It is designed to test the accuracy of the free surface representation in geodynamics code.

The model box spans 2800km by $700-1100$km (greater model height is necessary when employing sticky air on top). The initial condition is specified by a mantle of 600km thickness, overlain by a cosine shaped, $93-107$km-thick lithosphere:
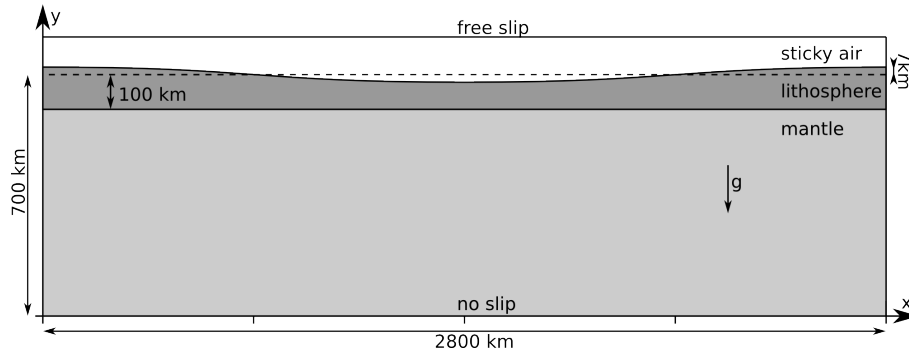
The sticky air layer has a thickness varying between 10 and 400km. The lithosphere is a highly viscous, dense medium ($\rho_L = 3300\text{kg m}^{-3}$, $\mu_L = 10^{23}\text{Pa s}$). The underlying ambient mantle has a density $\rho_M = 3300\text{kg m}^{-3}$ and a viscosity $\mu_M = 10^{21}\text{Pa s}$.
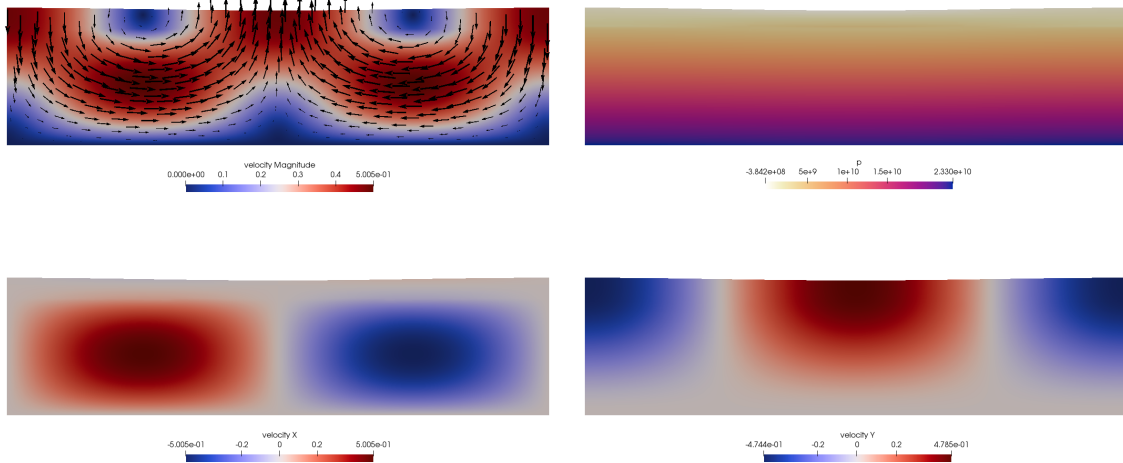
Results at $t = 0$ obtained wirh ASPECT by running the included cookbook.

The sticky air layer on the top has a density $\rho_{air} = 0\text{kg m}^{-3}$ and a viscosity $\mu_{air} = 10^{18}-10^{20}\text{Pa s}$ and is bordered by a free-slip top boundary condition. Free slip is also imposed at the sides while the bottom boundary is set to no slip condition.

The setup for the real free-surface model is identical to the setup described above, but the weak surface layer is removed and replaced by zero normal stress boundary conditions.

An analytical solution is presented by [1102]: the maximum topography at time t can be derived analytically using the relaxation rate γ and from the initial maximum topography h_{init}:

$$h_{analytic} = h_{init} \exp(\gamma t) \tag{20.106}$$

where $t = 14.825$kyr is the characteristic relaxation time and $\gamma = -0.2139 \times 10^{-11}$s^{-1} is the characteristic relaxation rate of the three-layer case at a given wavelength of 2800km. It should be noted that these values are valid for infinitesimal amplitudes, whereas deviations are to be expected for small but finite amplitudes. In particular, keeping the interface between the middle and lower layer flat and assuming a finite amplitude of the interface between the upper and middle layer implies that the thickness of the highly viscous middle layer varies laterally by $\pm 7\%$ (in the case of an initial maximum topography of 7km). This variation increases the effective viscous flexural rigidity and leads to a slightly longer relaxation time. The system is let to relax over time (typically 200kyrs) and the position of the free surface at $x = 0$ is recorded over time.
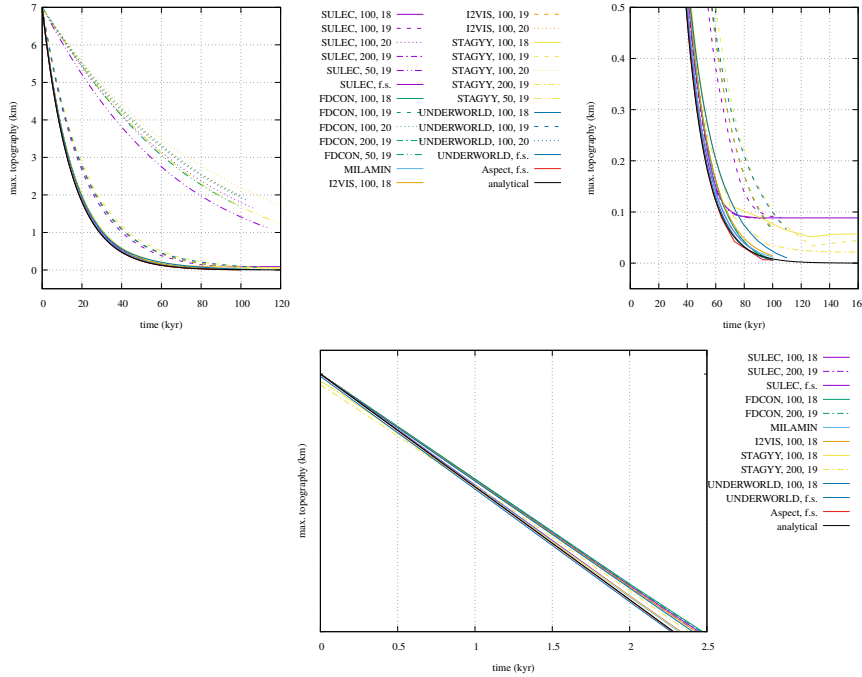
Data from the original paper. Aspect data are obtained by running the available example in the code.

20.27 3D spherical shell convection benchmark

The governing equations are those of an incompressible fluid with constant viscosity whose density depends on temperature. The fluid convects in a three-dimensional hollow sphere. The inner radius is 11/9 while the outer radius is 20/9, so that the depth of the mantle is exactly 1. Boundary conditions are free-slip at the top and bottom boundaries and isothermal with nondi- mensional temperatures of 0 and 1 at the top and bottom boundaries, respectively.

The dynamics of the system are gouverned by the Rayleigh number:

$$\mathsf{Ra} = \frac{\rho_0 \alpha g \ \Delta T \ \Delta R^3}{\kappa \eta}$$

Initial conditions for temperature are given as a function of coordinates with perturbations at some given spherical harmonics superimposed on a conductive temperature profile:

$$T(r,\theta,\phi) = \frac{R_i(r - R_o)}{r(R_i - R_o)} + \sum_m \left[\epsilon_{c,m}\cos m\phi + \epsilon_{s,m}\sin m\phi\right] p_{lm}(\theta)\sin\left(\pi\frac{r - R_i}{R_o - R_i}\right)$$

The first term represents a purely conductive temperature profile, while the second term is a perturbation to this profile, determining the final patterns of polyhedral symmetry. l and m are the spherical harmonic degree and order, respectively. $\epsilon_{c,m}$ and $\epsilon_{s,m}$ are the magnitudes of the individual spherical harmonic constituents. p_{lm} is a normalized associated Legendre polynomial that is related to the associated Legendre polynomial P_{lm} as[6]:

$$p_{lm}(\theta) = \sqrt{\frac{(2l+1)(l-m)!}{2\pi(1+\delta_{m0})(l+m)!}}P_l^m(\theta)$$

where P_l^m are the (unnormalized) associated Legendre functions and δ_{m0} is the Kronecker delta.

Note that there are somewhat subtle notation differences between the papers reporting on this benchmark with regards to the spherical harmonic parts and the normalisation term. Also note that the $(1+\delta_{m0})$ term is often omitted from spherical harmonics libraries (see ASPECT documentation).

In each case, we compute, as a function of time, Nusselt numbers for both the top and bottom boundaries, Nu_t and Nu_b, averaged temperature for the whole mantle, $<T>$ and averaged RMS velocity v_{rms} for the whole mantle with

$$Nu_t = \frac{R_o(R_o - R_i)}{R_i}Q_t \qquad Nu_b = \frac{R_i(R_o - R_i)}{R_o}Q_b$$

where Q_t and Q_b are the surface and bottom heat fluxes.
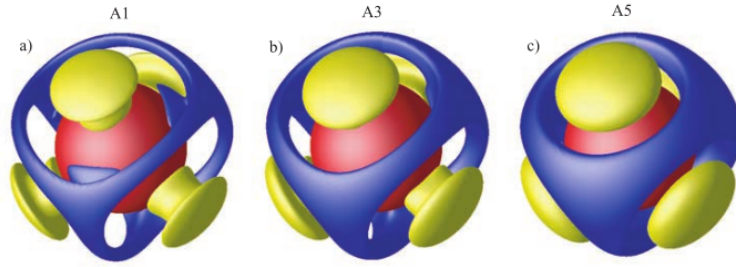
Taken from Zhong *et al.* (2008) [1521]. Representative steady state residual temperature $\delta T = T(r,\theta,\phi) - \langle T(r)\rangle$ for cases a, b, c.

Relevant Literature: Zhong *et al.* [1521] (CITCOMS), Arrial *et al.* [30] (CITCOMS vs. radial basis function), Shahnas *et al.* [1229] (own code), Liu & King [847] (ASPECT), Simon Bauer, Markus Huber, Siavash Ghelichkhan, Marcus Mohr, Ulrich Rüde, and B Wohlmuth. "Large-scale simulation of mantle convection based on a new matrix-free approach". In: *Journal of Computational Science* 31 (2019), pp. 60–76. DOI: 10.1016/j.jocs.2018.12.006 (2019).

20.28   2D convection benchmark ('Blankenbach *et al.* benchmark')

The abstract of the original publication by Blankenbach *et al.* (1989) [102] reads:

---

[6]There is a typo in [1229]

We have carried out a comparison study for a set of benchmark problems which are relevant for convection in the Earth's mantle. The cases comprise steady isoviscous convection, variable viscosity convection and time-dependent convection with internal heating. We compare Nusselt numbers, velocity, temperature, heat-flow , topography and geoid data. Among the applied codes are finite-difference, finite-element and spectral methods. In a synthesis we give best estimates of the 'true' solutions and ranges of uncertainty. We recommend these data for the validation of convection codes in the future.

The temperature is fixed to zero on top and to ΔT at the bottom, with reflecting symmetry at the sidewalls (i.e. $\partial_x T = 0$) and there are no internal heat sources. Free-slip conditions are implemented on all boundaries.

The Rayleigh number is given by

$$\mathsf{Ra} = \frac{\alpha g_y \Delta T h^3}{\kappa \nu} = \frac{\alpha g_y \Delta T h^3 \rho^2 c_p}{k \eta}$$

The initial temperature field is given by

$$T(x, y) = (1 - y) - 0.01 \cos(\pi x)\sin(\pi x)$$

The perturbation in the initial temperature fields leads to a perturbation of the density field and sets the fluid in motion.

Depending on the initial Rayleigh number, the system ultimately reaches a steady state after some time.
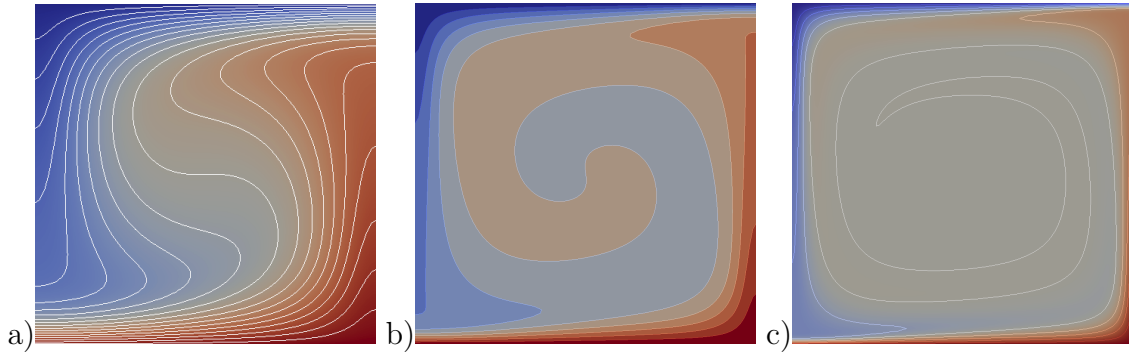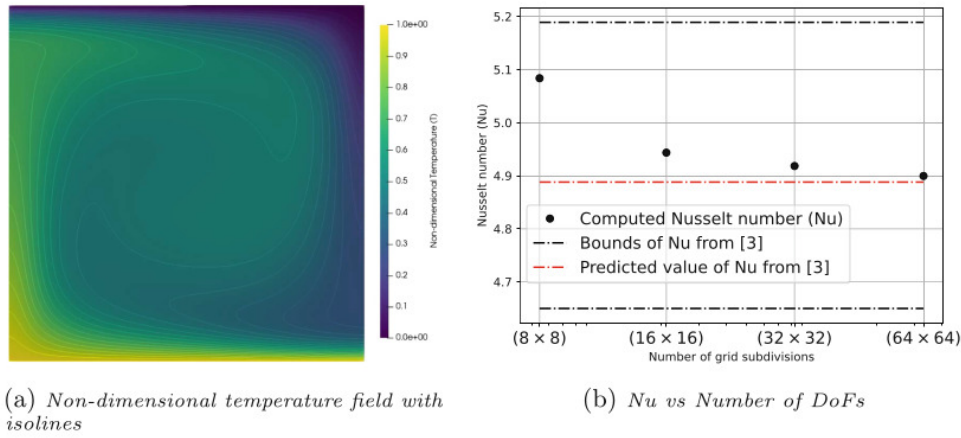

Temperature fields at steady-state for $\mathsf{Ra} = 10^4$ (a), $\mathsf{Ra} = 10^5$ (b), $\mathsf{Ra} = 10^6$ (c). Obtained with ELEFANT code [1342].

a)

Mesh size	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$	Reference
# DoFs	5 276	20 532	80 996	3.2×10^5	1.3×10^6	—
Period	0.048231	0.048051	0.048031	0.048030	0.048029	0.04803 ± 0.00003
Nu^{max}	7.4065	7.3822	7.3789	7.3788	7.3788	7.379 ± 0.005
Nu^{min}	6.5062	6.4717	6.4691	6.4691	6.4692	6.468 ± 0.005
Nu^{max}	7.2637	7.2047	7.1969	7.1960	7.1960	7.196 ± 0.005
Nu^{min}	6.7878	6.7949	6.7961	6.7965	6.7966	6.796 ± 0.005
v_{rms}^{max}	60.726	60.398	60.361	60.359	60.360	60.367 ± 0.015
v_{rms}^{min}	31.829	31.965	31.981	31.981	31.982	31.981 ± 0.02
v_{rms}^{max}	58.225	57.517	57.442	57.437	57.436	57.43 ± 0.05
v_{rms}^{min}	30.392	30.330	30.324	30.323	30.322	30.32 ± 0.03

b)

Finest mesh size	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$	Reference
# DoFs	$4.5 \ldots 6.0 \times 10^4$	$1.6 \ldots 2.2 \times 10^5$	$5.6 \ldots 8.0 \times 10^5$	—
Period	0.048029	0.048030	0.048030	0.04803 ± 0.00003
Nu^{max}	7.3809	7.3792	7.3788	7.379 ± 0.005
Nu^{min}	6.4718	6.4695	6.4691	6.468 ± 0.005
Nu^{max}	7.1996	7.1967	7.1960	7.196 ± 0.005
Nu^{min}	6.7986	6.7969	6.7965	6.796 ± 0.005
v_{rms}^{max}	60.366	60.361	60.360	60.367 ± 0.015
v_{rms}^{min}	31.980	31.981	31.981	31.981 ± 0.02
v_{rms}^{max}	57.449	57.434	57.435	57.43 ± 0.05
v_{rms}^{min}	30.322	30.322	30.322	30.32 ± 0.03

In Ilangovan, D'Ascoli, Kohl, and Mohr [652] (2024) we find:

(a) *Non-dimensional temperature field with isolines*

(b) *Nu vs Number of DoFs*

		v_{rms}	Nu
Blankenbach *et al.* (1989) [102]	$\mathtt{Ra} = 10^4$	42.864947 ± 0.000020	4.884409 ± 0.000010
	$\mathtt{Ra} = 10^5$	193.21454 ± 0.00010	10.534095 ± 0.000010
	$\mathtt{Ra} = 10^6$	833.98977 ± 0.00020	21.972465 ± 0.000020
Tackley (1994) [1311]	$\mathtt{Ra} = 10^4$	42.775	4.878
	$\mathtt{Ra} = 10^5$	193.11	10.531
	$\mathtt{Ra} = 10^6$	833.55	21.998
King (2009) [743]	$\mathtt{Ra} = 10^4$	42.867	4.885
	$\mathtt{Ra} = 10^5$	193.248	10.536
	$\mathtt{Ra} = 10^6$	834.353	21.981
Thieulot (2014) [1342]	$\mathtt{Ra} = 10^4$	42.867	4.882
	$\mathtt{Ra} = 10^5$	193.255	10.507
	$\mathtt{Ra} = 10^6$	834.712	21.695
Aspect [48]	$\mathtt{Ra} = 10^4$		
	$\mathtt{Ra} = 10^5$		
	$\mathtt{Ra} = 10^6$		

Steady state Nusselt number and Vrms measurements as reported in the literature and obtained with Elefant on a 200×200 grid. King (2009) results on 200x200 grid with ConMan.

🛢 Relevant Literature: Travis et al. [1368] (1990), Ogawa [1010] (1993), Trompert and Hansen [1373] (1998), Auth and Harder [35] (1999), Christon, Gresho, and Sutton [274] (2002), Chiu-Webster, Hinch, and Lister [252] (2008), Kameyama, Kageyama, and Sato [706] (2005), King [743] (2009), Beuchert and Podladchikov [92] (2010), Leng and Zhong [815] (2011), Davies, Wilson, and Kramer [328] (2011), Vynnytska, Rognes, and Clark [1432] (2013), Trim, Butler, and Spiteri [1372] (2021), Davies, Kramer, Ghelichkhan, and Gibson [325] (2002), Sime and Wilson [1247] (2020), Ilangovan, D'Ascoli, Kohl, and Mohr [652] (2024), stone 3.

20.29 Subduction 'benchmark' of Schmeling *et al.* (2008)

The setup is as follows :
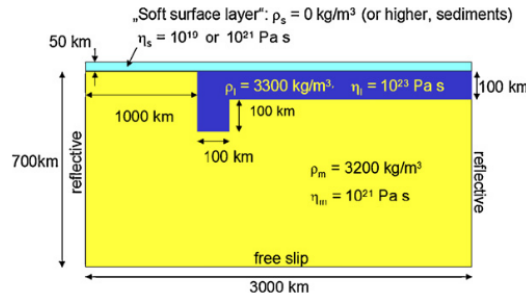
Taken from Schmeling *et al.* (2008) [1200].

Materials are linear viscous, initial geometry is simple, boundary conditions are simple. On paper it sounds like a good idea. See STONE 67 for a discussion on why this setup was doomed from the beginning as a benchmark.

Experiments have been conducted by a handful of codes, investigating the effect of averaging and mesh resolution:
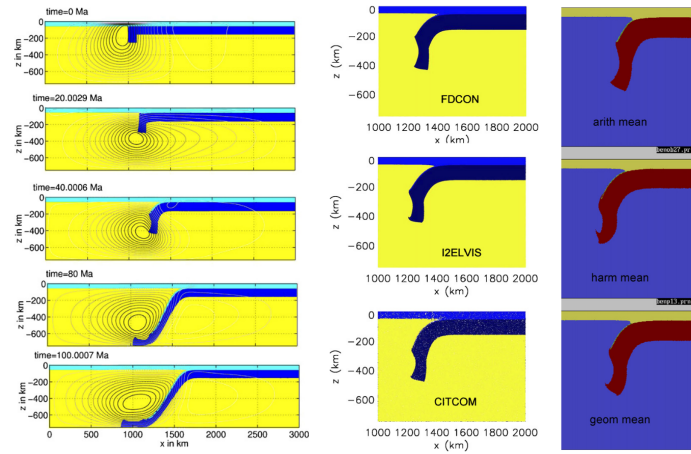
Taken from Schmeling *et al.* (2008) [1200]. Left: case 1 model (here FDCON-4 is shown). Streamlines are also shown. Middle: Shapes of different case 1 models at similar stages: FDCON: 40 Myears, I2ELVIS: 34.7 Myears, CITCOM: 38.1 Myears. Viscosity averaging: geometric mean in all cases. Right: Comparison of the shapes of the slabs for different viscosity averaging meth- ods using I2VIS. Note that the snapshots are taken at different times (59.6, 24.4, 37.8 Myears from top to bottom), so that the slab tips have reached comparable levels.
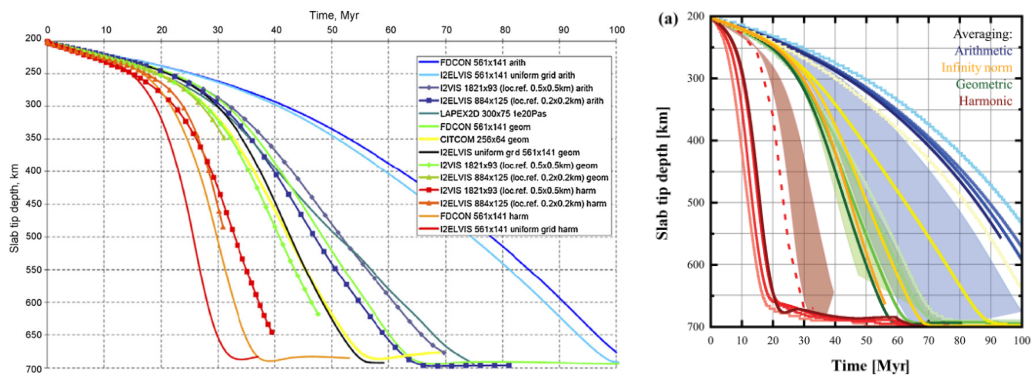
Temporal behaviour of case 1 modelled by different codes with highest resolutions each. Each curve shows the position of the deepest part of the slab (slab tip) as a function of time below the initial surface of the lithosphere. Left: taken from Schmeling *et al.* (2008) [1200]; Right:taken from Glerum *et al.* (2018) [496].

1117
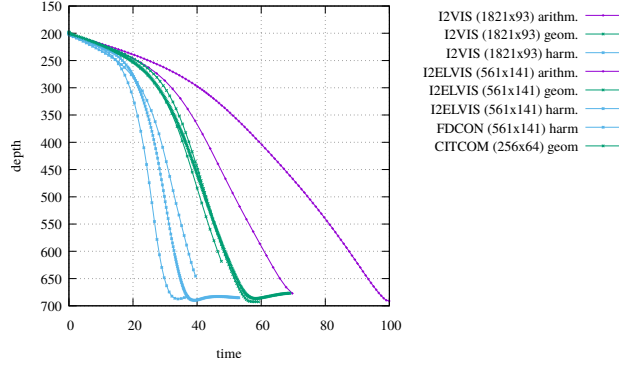
Data curtesy of Prof. H. Schmeling.

20.30 Thin layer entrainment

Data pertaining to this section are to be found at:
https://github.com/cedrict/fieldstone/tree/master/images/benchmark_thinlayer

The problem is a simulation to study the amount of entrainment by thermal convection of a dense, thin layer at the bottom of the model [1406]. To the author's knowledge only two other publications (Tackley & King (2003) [1313], van Thienen (2007) [1410]) have presented results pertaining to this benchmark. The results shown here after are obtained with my ELEFANT code using the particle-in-cell technique and originate in the ELEFANT paper [1342].

The box is 2×1, and contains two fluids:
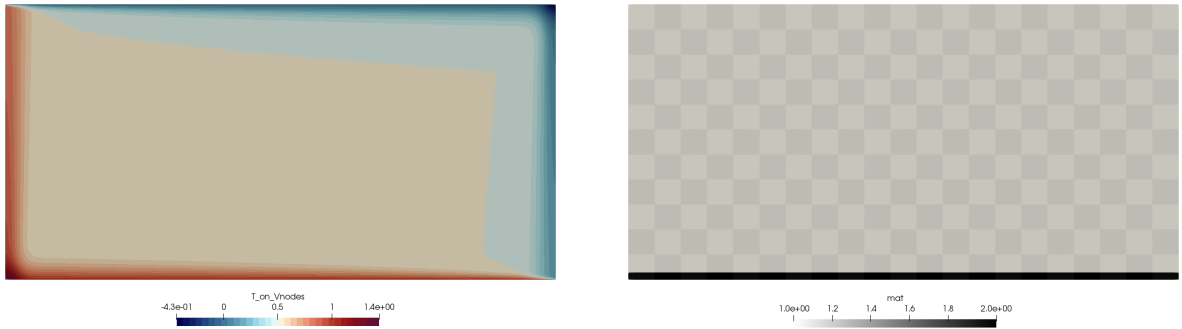
Fluid 1 has a density $\rho_1 = 1$ and a viscosity $\eta = 1$. Fluid 2 is heavier ($\rho_2 = \rho_1 + \Delta\rho$) but has the same viscosity. Both fluids have a thermal expansion coefficient $\alpha = 10^{-10}$, a thermal conductivity $k = 1$, and a heat capacity coefficient $c_p = 1$. Fluid 2 is placed at the bottom of the box ($0 \leq y \leq 0.025$).

This experiment is parameterised by the thermal Rayleigh number $\mathsf{Ra} = 300,000$ and and the compositional Rayleigh number $\mathsf{Ra}_c = 450,000$ which are defined as follows:

$$\mathsf{Ra}_T = \frac{\alpha\rho g\Delta T L_y^3}{\kappa\eta} = \frac{\alpha\rho^2 g\Delta T L_y^3 c_p}{k\eta} = \alpha g \tag{20.107}$$

$$\mathsf{Ra}_c = \frac{\Delta\rho g L_y^3}{\kappa\eta} = \frac{\rho\Delta\rho g L_y^3 c_p}{k\eta} = \Delta\rho g \tag{20.108}$$

1118

where I have used the relationship $\kappa = k/\rho c_p$. B is defined as $B = \mathsf{Ra}_T/\mathsf{Ra}_c$ so The gravity acceleration is therefore set to $g = \mathsf{Ra}/\alpha$ and this yields $\Delta\rho = \mathsf{Ra}_c/g = B\mathsf{Ra}_T/g = B \times \alpha$.

Free-slip boundary conditions are imposed on all sides of the domain. Temperature boundary conditions are $T(x, y = 0) = 1$ and $T(x, y = 1) = 0$. The analytical initial temperature field is given by

$$T(x, y) = T_u(x, y) + T_l(x, y) + T_r(x, y) + T_s(x, y) - \frac{3}{2} \tag{20.109}$$

where

$$
\begin{aligned}
T_u(x, y) &= \frac{1}{2}\text{erf}\left(\frac{1-y}{2}\sqrt{\frac{u_0}{x}}\right) \\
T_l(x, y) &= 1 - \frac{1}{2}\text{erf}\left(\frac{y}{2}\sqrt{\frac{u_0}{L_x - x}}\right) \\
T_r(x, y) &= \frac{1}{2} + \frac{Q}{2\sqrt{\pi}}\sqrt{\frac{u_0}{y+1}}\exp\left(-\frac{x^2 u_0}{4y + 4}\right) \\
T_s(x, y) &= \frac{1}{2} - \frac{Q}{2\sqrt{\pi}}\sqrt{\frac{u_0}{2-y}}\exp\left(-\frac{(L_x - x)^2 u_0}{8 - 4y}\right)
\end{aligned}
\tag{20.110}
$$

with

$$u_0 = \frac{L_x^{7/3}}{(1 + L_x^4)^{2/3}}\left(\frac{\mathsf{Ra}}{2\sqrt{\pi}}\right)^{2/3} \qquad Q = 2\sqrt{\frac{L_x}{\pi u_0}} \tag{20.111}$$

Using $L_x = 2$, $\mathsf{Ra} = 3 \times 10^5$, one gets $u_0 \simeq 1469.315$ and $Q \simeq 0.0416305$.

Given the small thickness of the bottom layer, it seems quite legitimate to investigate the influence of grid resolution on the simulation. I have therefore looked at the initial root mean square velocity measurement as a function of the element diagonal value (a proxy for the average resolution in the case where elements are not square).

Results are confirm that the element size plays a non negligible role at startup on the dynamics of the system. Superimposed on the figure are the measurements provided by Prof. van Keken (black squares in the gray box). They agree well with my measurements but also indicate that none of the authors in the original study ran the experiment at a high-enough resolution to start with (their results were therefore most likely resolution dependent).

We see that the number of markers per element at startup is critical at (very) low resolution but that it does not lead to significant velocity variations at high resolution.
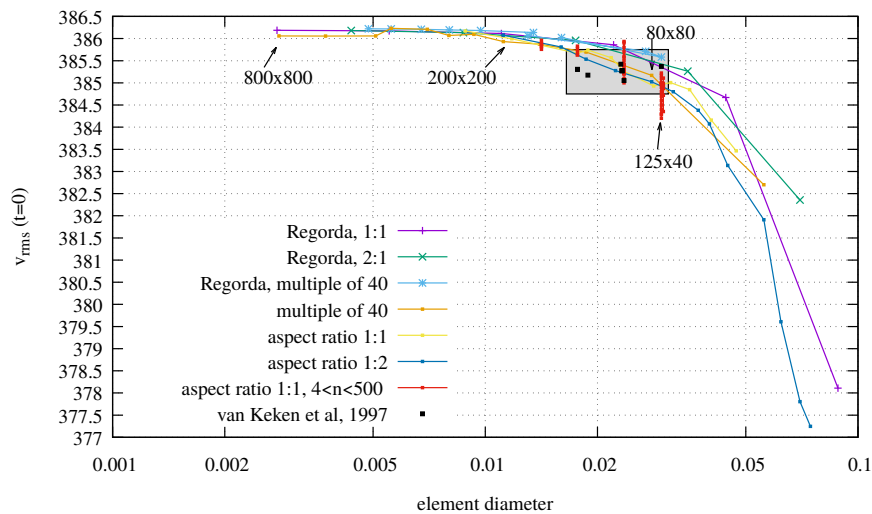
Thin layer entrainment experiment: root mean square velocity measurements at $t = 0$ as a function of the element diagonal size. The red square points correspond to resolutions where the number of elements in each direction is a multiple of 40 (i.e. L_y/d), so that no element would contain a mix of fluids 1

1119

Looking at the root mean square velocity measurements, we see that the measurements done with Elefant agree nicely with those presented in van Keken *et al.* [1406]. Past $t \sim 0.015$, the curves diverge clearly across all codes and authors, so I only need to focus the comparison for times $t < 0.015$. For the three tested resolutions,measurements agree well and fall within the grey curves representing all results of van Keken *et al.* . Additional tests have been carried out concerning the value of the Courant number (0.1 to 0.25) and the initial number of markers per element (100 or 200) and these parameters led to extremely similar results.
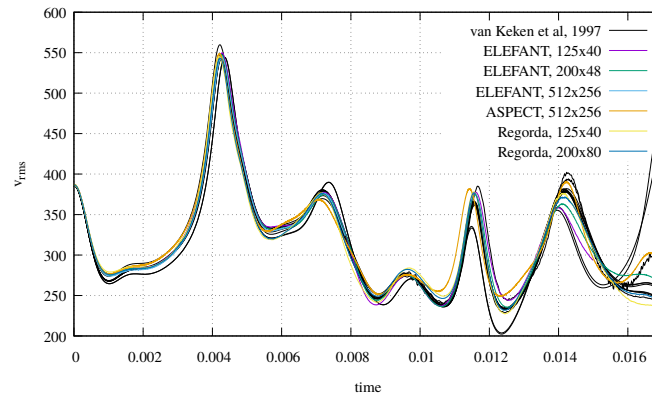
Thin layer entrainment experiment. Root mean square velocity as a function of time. All results presented in van Keken *et al.* (1997) are collapsed in black

dashed lines. All simulations were run with an initial marker density of 100 markers per element and with a Courant numberof 0.25. Taken from

Elefant paper [1342].

As observed in van Keken *et al.* , the dense layer is first swept into the lower left corner. Thermal instabilities then further develop in an asymmetrical way and entrain the dense material. Past $t \simeq 0.015$ the system becomes more and more chaotic with markers being randomly mixed in the system in a non-orderly fashion.
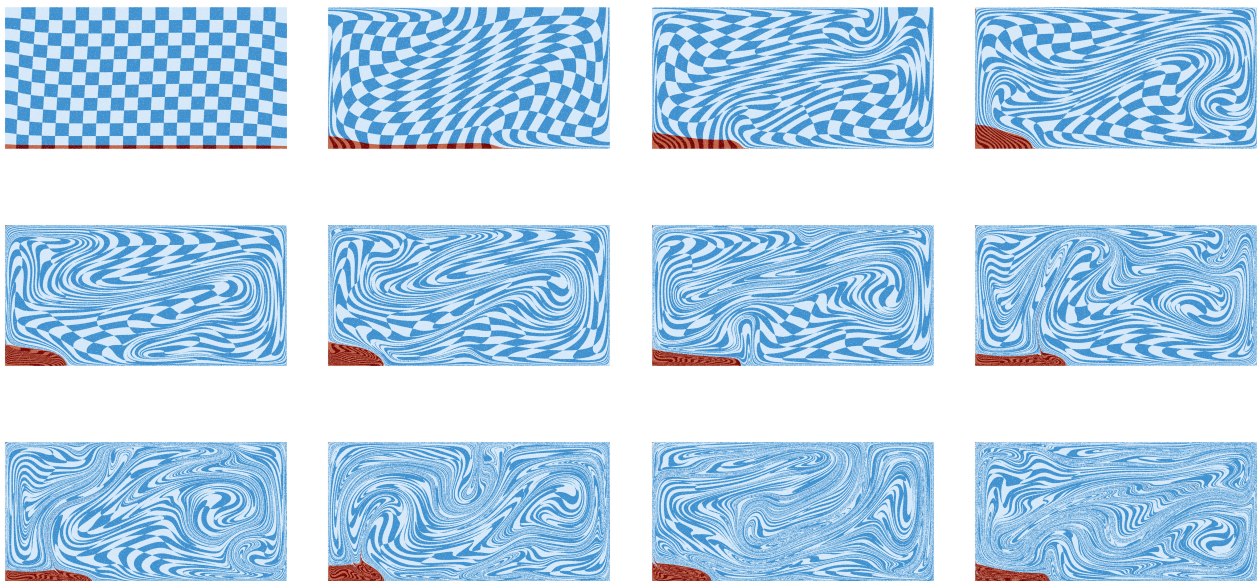
Marker distribution as obtained with Elefant for grid 240x120, init_marker_density=7,

random distribution, CFL=0.25, rkmethod=2, m_to_q=2. (unpublished).

🔷 Relevant Literature: Trim *et al.* (2020) [1371].

20.31 Cylindrical ring subject to a line load

This benchmark is found in Ergatoudis, Irons, and Zienkiewicz [402] (1968). Unfortunately there are no rheological parameters provided in the paper.
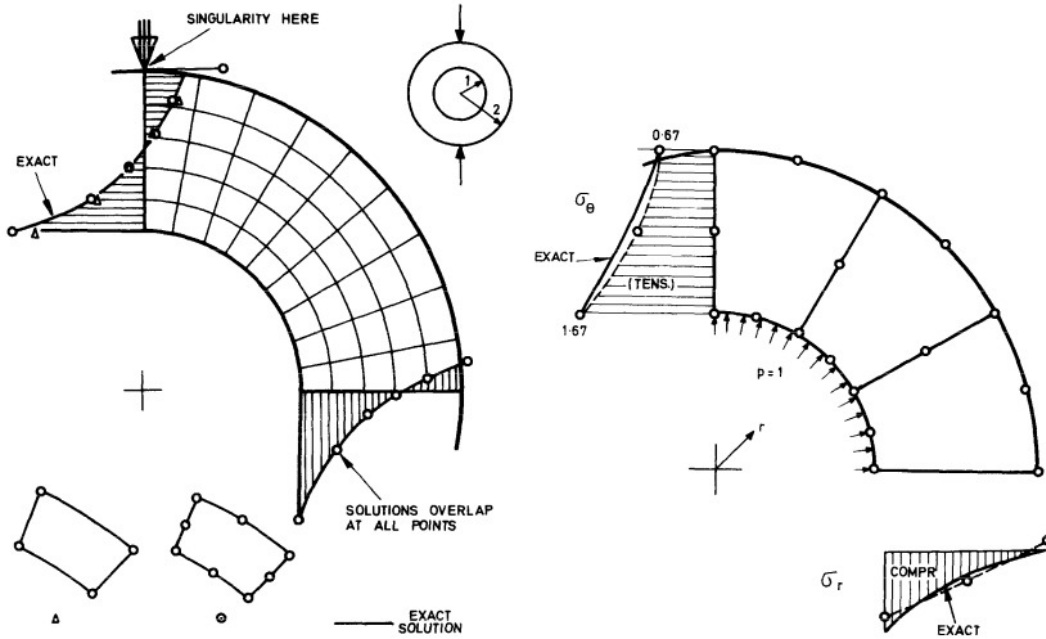
20.32 Lid-driven cavity in 3D

It is found in Bertrand, Gadbois, and Tanguy [89] (1992). Note that simulations were carried out at a Reynolds number of 400.

The motion of the fluid is induced by moving one of the walls of a unit cube at a constant velocity while the other walls remain stationary:
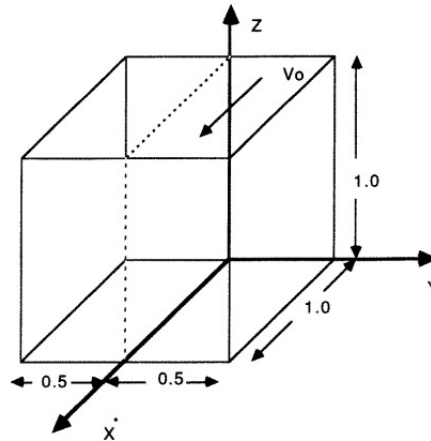
The boundary conditions are imposed as follows:

- on the moving wall: $u = u_0$, $v = w = 0$

- on the symmetry plane ($y = 0.5$): $v = 0$

- on the other walls: $u = v = w = 0$

1121

The lid-driven cavity problem is well suited for testing elements because of the wealth of numerical results available.

In order to compare the computed velocity fields with published data, the usual graph of the horizontal velocity profile $uvsz$ at the plane of symmetry was plotted.
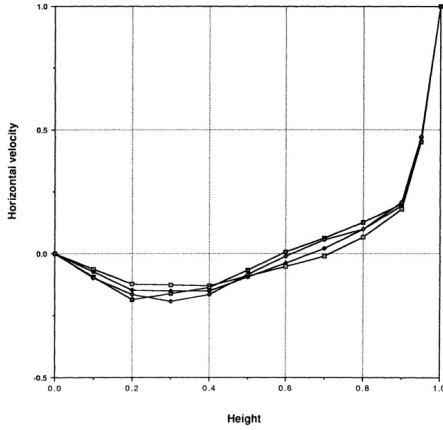
Figure 5. Comparison of horizontal velocity profiles: dotted-square = $P_1^+-P_0$; black diamond = $P_1^{++}-P_1$; white square = $P_2^+-P_1$; white diamond = reference solution
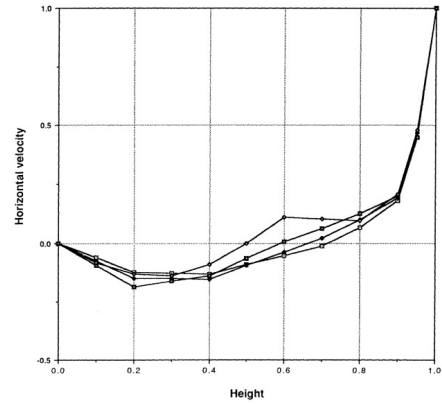
Figure 6. Comparison of horizontal velocity profiles: dotted-square = $P_1^+-P_0$; black diamond = $P_1^{++}-P_1$; white square = $P_2^+-P_1$; white diamond = Reddy' solution (1982)

Left: results obtained with various elements against the solution obtained by Robichaud and Tanguy [1152]; Right: results obtained with various elements against the solution obtained by Reddy [1117].

20.33 Hollow sphere under internal pressure

This benchmark was inspired by the one at XXX. A homogenous hollow sphere (internal radius a, external radius b) is made of an elastic perfectly plastic material (characterised by the Young's modulus E, the Poisson ratio ν and the yield stress σ_y). The yield criterion in von Mises. Starting from an initial stress-free state the internal pressure rises from 0 to p.
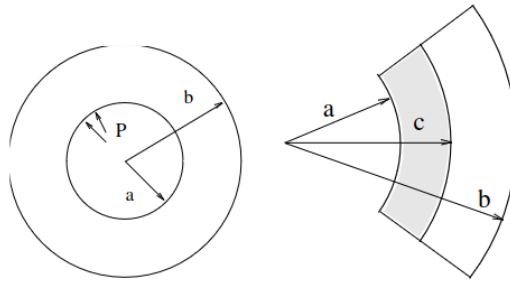
Left: geometry and applied load. Right: illustration of the plastic zone growing from the internal surface.

Since the body's geometry and the loading present a spherical symmetry, and since the material behaviour is isotropic the solution is formulated in spherical coordinates (r, θ, ϕ). The displacement field is then given by

$$u_r = f(r) \tag{20.112}$$
$$u_\theta = 0 \tag{20.113}$$
$$u_\phi = 0 \tag{20.114}$$

1122

The strain tensor in spherical coordinates is given by

$$
\begin{aligned}
\varepsilon_{rr} &= \frac{\partial u_r}{\partial r} \\
\varepsilon_{\theta\theta} &= \frac{u_r}{r} + \frac{1}{r}\frac{\partial u_\theta}{\partial \theta} \\
\varepsilon_{\phi\phi} &= \frac{1}{r\sin\theta}\frac{\partial u_\phi}{\partial \phi} + \frac{u_r}{r} + \frac{u_\theta \cot\theta}{r} \\
\varepsilon_{\theta r} = \varepsilon_{r\theta} &= \frac{1}{2}\left(r\frac{\partial}{\partial r}(\frac{u_\theta}{r}) + \frac{1}{r}\frac{\partial u_r}{\partial \theta}\right) \\
\varepsilon_{\phi r} = \varepsilon_{r\phi} &= \frac{1}{2}\left(\frac{1}{r\sin\theta}\frac{\partial u_r}{\partial \phi} + r\frac{\partial}{\partial r}(\frac{u_\phi}{r})\right) \\
\varepsilon_{\phi\theta} = \varepsilon_{\theta\phi} &= \frac{1}{2}\left(\frac{\sin\theta}{r}\frac{\partial}{\partial \theta}(\frac{u_\phi}{\sin\theta}) + \frac{1}{r\sin\theta}\frac{\partial u_\theta}{\partial \phi}\right)
\end{aligned}
$$

which, given Eqs. (20.112),(20.113),(20.114), leads to:

$$
\begin{aligned}
\varepsilon_{rr} &= \frac{\partial u_r}{\partial r} \\
\varepsilon_{\theta\theta} &= \frac{u_r}{r} \\
\varepsilon_{\phi\phi} &= \frac{u_r}{r} \\
\varepsilon_{\theta r} = \varepsilon_{r\theta} &= 0 \\
\varepsilon_{\phi r} = \varepsilon_{r\phi} &= 0 \\
\varepsilon_{\phi\theta} = \varepsilon_{\theta\phi} &= 0
\end{aligned}
$$

because we also have $\partial_\theta \to 0$ and $\partial_\phi \to 0$. Then we can compute a quantity that we will need later:

$$
\vec{\nabla} \cdot \vec{u} = \frac{\partial u_r}{\partial r} + 2\frac{u_r}{r}
$$

Finally the stress tensor is given by

$$
\boldsymbol{\sigma} = \lambda(\vec{\nabla} \cdot \vec{u})\mathbf{1} + 2\mu\varepsilon(\vec{u})
$$

or, more explicitely:

$$
\begin{aligned}
\sigma_{rr} &= \lambda\left(\frac{\partial u_r}{\partial r} + 2\frac{u_r}{r}\right) + 2\mu\frac{\partial u_r}{\partial r} \\
\sigma_{\theta\theta} &= \lambda\left(\frac{\partial u_r}{\partial r} + 2\frac{u_r}{r}\right) + 2\mu\frac{u_r}{r} \\
\sigma_{\phi\phi} &= \lambda\left(\frac{\partial u_r}{\partial r} + 2\frac{u_r}{r}\right) + 2\mu\frac{u_r}{r} \\
\sigma_{r\theta} &= 2\mu\varepsilon_{r\theta} = 0 \\
\sigma_{r\phi} &= 2\mu\varepsilon_{r\phi} = 0 \\
\sigma_{\theta\phi} &= 2\mu\varepsilon_{\theta\phi} = 0
\end{aligned}
\tag{20.115}
$$

The strong form of the PDE that governs force balance in a medium is given by

$$
\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{f} = \vec{0}
$$

1123

In this case $\vec{f} = \vec{0}$ (no body force). The divergence of the stress tensor in spherical coordinates is given by[7]:

$$
\begin{aligned}
\frac{\partial \sigma_{rr}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{r\theta}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \sigma_{r\phi}}{\partial \phi} + \frac{1}{r}(2\sigma_{rr} - \sigma_{\theta\theta} - \sigma_{\phi\phi} + \sigma_{r\theta}\cot\theta) &= 0 \\
\frac{\partial \sigma_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{\theta\theta}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \sigma_{\theta\phi}}{\partial \phi} + \frac{1}{r}\left((\sigma_{\theta\theta} - \sigma_{\phi\phi})\cot\theta + 3\sigma_{r\theta}\right) &= 0 \\
\frac{\partial \sigma_{r\phi}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{\theta\phi}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial \sigma_{\phi\phi}}{\partial \phi} + \frac{1}{r}\left(2\sigma_{\theta\phi}\cot\theta + 3\sigma_{r\phi}\right) &= 0
\end{aligned}
$$

Using the equations above for the stress components, using $\partial_\theta \to 0$ and $\partial_\phi \to 0$ and remembering that $\sigma_{\theta\theta} = \sigma_{\phi\phi}$ we find that the lhs of the second and third lines are identically zero, and the first line becomes simply

$$
\frac{\partial \sigma_{rr}}{\partial r} + \frac{1}{r}(2\sigma_{rr} - \sigma_{\theta\theta} - \sigma_{\phi\phi}) = 0
$$

or,

$$
\frac{d\sigma_{rr}}{dr} + \frac{2}{r}(\sigma_{rr} - \sigma_{\theta\theta}) = 0 \tag{20.116}
$$

The elastic constitutive equations can also be written[8]:

$$
\varepsilon_{ij} = \frac{1}{2\mu}\sigma_{ij} - \frac{\nu}{E}\delta_{ij}\sigma_{kk} = \frac{1}{E}\left[(1+\nu)\sigma_{ij} - \nu\delta_{ij}\sigma_{kk}\right]
$$

We have $\sigma_{kk} = \sigma_{rr} + \sigma_{\theta\theta} + \sigma_{\phi\phi} = \sigma_{rr} + 2\sigma_{\theta\theta}$ so that explicitely

$$
\begin{aligned}
E\varepsilon_{rr} &= (1+\nu)\sigma_{rr} - \nu(\sigma_{rr} + 2\sigma_{\theta\theta}) \\
E\varepsilon_{\theta\theta} &= (1+\nu)\sigma_{\theta\theta} - \nu(\sigma_{rr} + 2\sigma_{\theta\theta}) \\
E\varepsilon_{\phi\phi} &= (1+\nu)\sigma_{\phi\phi} - \nu(\sigma_{rr} + 2\sigma_{\theta\theta})
\end{aligned}
$$

or (removing the useless third line),

$$
\begin{aligned}
E\varepsilon_{rr} &= \sigma_{rr} - 2\nu\sigma_{\theta\theta} \\
E\varepsilon_{\theta\theta} &= \sigma_{\theta\theta}(1-\nu) - \nu\sigma_{rr}
\end{aligned} \tag{20.117}
$$

After replacing strain by its expression as a function of displacement, it comes (using $\lambda = E\nu/(1 - 2\nu)/(1+\nu)$):

$$
\begin{aligned}
\sigma_{rr} &= \frac{\lambda}{\nu}\left((1-\nu)\frac{du_r}{dr} + 2\nu\frac{u_r}{r}\right) \\
\sigma_{\theta\theta} &= \frac{\lambda}{\nu}\left(\nu\frac{du_r}{dt} + \frac{u_r}{r}\right)
\end{aligned} \tag{20.118}
$$

These two equations can be introduced in the equilibrium equation (20.116), leading to :

$$
\frac{d^2 u_r}{dr^2} + \frac{2}{r}\frac{du_r}{dr} - \frac{2}{r^2}u_r = 0
$$

The solution of this equation is :

$$
u_r(r) = C_1 r + \frac{C_2}{r^2}
$$

---

[7]https://en.wikipedia.org/wiki/Linear_elasticity
[8]https://en.wikipedia.org/wiki/Linear_elasticity

By replacing u_r in the preceding expressions, it comes:

$$\sigma_{rr} = \frac{\lambda}{\nu}\left((1+\nu)C_1 - 2(1-2\nu)\frac{C_2}{r^3}\right)$$

$$\sigma_{\theta\theta} = \frac{\lambda}{\nu}\left((1+\nu)C_1 + (1-2\nu)\frac{C_2}{r^3}\right)$$

The integration constants C_1 and C_2 can be determined from the boundary conditions[9]:

$$\sigma_{rr}(r=b) = 0 \quad \Rightarrow \quad C_2 = \frac{1+\nu}{2(1-2\nu)}b^3 C_1$$

$$\sigma_{rr}(r=a) = -p \quad \Rightarrow \quad C_1 = \frac{1-2\nu}{E}\frac{a^3}{b^3-a^3}p$$

so that in the end:

$$\sigma_{rr} = -\frac{a^3}{b^3-a^3}\left(\frac{b^3}{r^3}-1\right)p$$

$$\sigma_{\theta\theta} = \sigma_{\phi\phi} = \frac{a^3}{b^3-a^3}\left(\frac{b^3}{2r^3}+1\right)p$$

$$u_r = \frac{a^3}{b^3-a^3}\left((1-2\nu)r + (1+\nu)\frac{b^3}{2r^2}\right)\frac{p}{E} \qquad (20.119)$$

Onset of plastic flow   Let us recall Eq. (4.75) for a diagonal stress tensor:

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6}\left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2\right]$$

then we have

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6}\left[(\sigma_{rr}-\sigma_{\theta\theta})^2 + (\sigma_{\theta\theta}-\sigma_{\phi\phi})^2 + (\sigma_{rr}-\sigma_{\phi\phi})^2\right] = \frac{1}{6}\left[(\sigma_{rr}-\sigma_{\theta\theta})^2 + (\sigma_{rr}-\sigma_{\theta\theta})^2\right] = \frac{1}{3}(\sigma_{rr}-\sigma_{\theta\theta})^2$$

or,

$$\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} = \frac{1}{\sqrt{3}}|\sigma_{rr} - \sigma_{\theta\theta}|$$

and the von Mises criterion writes (see Eq. (5.21)):

$$\sqrt{3}\sqrt{\mathcal{I}_2(\boldsymbol{\tau})} - \sigma_y = 0$$

or,

$$|\sigma_{rr} - \sigma_{\theta\theta}| - \sigma_y = 0$$

Let us compute the stress term:

$$\sigma_{rr} - \sigma_{\theta\theta} = -\frac{a^3}{b^3-a^3}\left(\frac{b^3}{r^3}-1\right)p - \frac{a^3}{b^3-a^3}\left(\frac{b^3}{2r^3}+1\right)p$$

$$= -\frac{a^3}{b^3-a^3}\frac{3}{2}\frac{b^3}{r^3}p \qquad (20.120)$$

Since $b > a$ and $p > 0$ then

$$|\sigma_{rr} - \sigma_{\theta\theta}| = \frac{a^3}{b^3-a^3}\frac{3}{2}\frac{b^3}{r^3}p$$

The plastic yield is reached when $|\sigma_{\theta\theta} - \sigma_{rr}|$, increasing function of p, is equal to σ_y, the yield limit in onedimensional tension. The initial plastic zone is then located in $r = a$, and the value of the pressure Pe is:

$$P_e = \frac{2}{3}\left(1 - \frac{a^3}{b^3}\right)\sigma_y$$

---

[9]I have not 100% checked these expressions

Chapter 21

Vorticity-stream function approach

The Stream function (commonly denoted by Ψ) approach is a useful approach in fluid dynamics as it can provide relatively quick solutions to 2D incompressible flow problems. Using a stream function formulation is numerically convenient because velocity information is contained in a single scalar equation and pressure vanishes from the solution process.

The stream function is a function of coordinates and time of an inviscid liquid. It allows to determine the components of velocity by differentiating the stream function with respect to the space coordinates. A family of curves $\Psi = constant$ represent streamlines, i.e. the stream function remains constant along a streamline. Although also valid in 3D, this approach is mostly used in 2D because of its relative simplicity.

Glaisner and Tezduyar [493] state:

> "The main advantages of the vorticity stream-function formulation are the simple form of the equations in two-dimensions and the in-built satisfaction of the incompressibility constraint.
>
> In two dimensions, the vorticity transport equation and the Poisson equation for the stream function are scalar and there are only two degree of freedom in the problem. Moreover, the vorticity stream-function form of the Navier-Stokes equations allows equal order of interpolation for the vorticity and the stream function. In fact the bilinear interpolations which have been used here for both the unknowns are sufficient which is an asset from the point of view of implementation.
>
> On the other hand, a flow field obtained by the solution of the vorticity stream-function equations is by definition divergence-free and the initial condition of the problem do not need to satisfy the Incompressibility constraint. This allows the use of an initial flow field noncontinuous at the boundary of the domain. "

Stream function	$\leftrightarrow$	Stroomfunctie
Vorticity	$\leftrightarrow$	vorticiteit

21.1 Vorticity-stream function formulation of the isoviscous Navier-Stokes equation

What follows is adapted from Glaisner and Tezduyar [493] (1987). The vorticity transport equation can be obtained by taking the curl of the (isoviscous) incompressible momentum equation,

$$\vec{\nabla} \times \left[\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} \right] = \vec{\nabla} \times \left[-\frac{1}{\rho}\vec{\nabla}p + \nu\vec{\nabla}^2\vec{v} + \vec{g} \right] \tag{21.1}$$

with $\nu = \eta/\rho$. i need to propagate gravity in what follows Computing term-by-term, we have

$$\vec{\nabla} \times \frac{\partial \vec{v}}{\partial t} = \frac{\partial}{\partial t}(\vec{\nabla} \times \vec{v}) = \frac{\partial \vec{\omega}}{\partial t}$$

where the vorticity vector $\vec{\omega}$ is defined as

$$\vec{\omega} = \vec{\nabla} \times \vec{v} = \begin{vmatrix} \partial_x & u \\ \partial_y & v \\ \partial_z & w \end{vmatrix} \tag{21.2}$$

Because the curl of a gradient is zero (assuming rho constant?!),

$$\vec{\nabla} \times \left(-\frac{1}{\rho}\vec{\nabla}p \right) = 0$$

Also, (assuming nu constant?!)

$$\vec{\nabla} \times (\nu\vec{\nabla}|\vec{v}|^2) = \nu\vec{\nabla}^2(\vec{\nabla} \times \vec{v}) = \nu\vec{\nabla}^2\vec{\omega}$$

Now, we use the vector identity

$$(\vec{v} \cdot \vec{\nabla})\vec{v} = \frac{1}{2}\vec{\nabla}|\vec{v}|^2 - \vec{v} \times (\vec{\nabla} \times \vec{v}) = \frac{1}{2}\vec{\nabla}|\vec{v}|^2 - \vec{v} \times \vec{\omega} \tag{21.3}$$

Taking the curl of both sides and making use of the curl of a gradient equals zero and $\vec{\omega} = \vec{\nabla} \times \vec{v}$, results in

$$\vec{\nabla} \times [(\vec{v} \cdot \vec{\nabla})\vec{v}] = -\vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times (\vec{\omega} \times \vec{v})$$

Combining all the above terms, we have thus obtained the vorticity equation

$$\boxed{\frac{\partial \vec{\omega}}{\partial t} + \vec{\nabla} \times (\vec{\omega} \times \vec{v}) = \nu\vec{\nabla}^2\vec{\omega}} \tag{21.4}$$

Also, we have the identity

$$\vec{\nabla} \times (\vec{\omega} \times \vec{v}) = (\vec{v} \cdot \vec{\nabla})\vec{\omega} - (\vec{\omega} \cdot \vec{\nabla})\vec{v}$$

so that in the end:

$$\boxed{\frac{\partial \vec{\omega}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{\omega} = (\vec{\omega} \cdot \vec{\nabla})\vec{v} + \nu\vec{\nabla}^2\vec{\omega}} \tag{21.5}$$

In the case of a two-dimensional flow, Eq. (21.5) simplifies to[1]

$$\frac{\partial \vec{\omega}}{\partial t} + \vec{v} \cdot \vec{\nabla}\vec{\omega} = \nu\vec{\nabla}^2\vec{\omega}$$

The velocity $\vec{v}$ being solenoidal, there exists a vector field $\vec{\Psi} = (0, 0, \Psi)$ such that

$$\vec{v} = \vec{\nabla} \times \vec{\Psi}$$

The nonzero component of Ψ is called the stream function. A relation between the vorticity and the stream function is obtained as

$$\vec{\omega} = \vec{\nabla} \times \vec{\nabla} \times \vec{\Psi} = -\vec{\nabla}^2\vec{\Psi}$$

---

[1]In the case of a two-dimensional flow, the vorticity-stretching term $(\vec{v} \cdot \vec{\nabla})\vec{v}$ is zero and the vorticity transport equation contains only one nonlinear term. This, together with the scalar nature of the equations, is the reason why the vorticity stream-function formulation is very convenient in two dimensions.

For 2D flows in the $x - y$ plane, since $w = 0$ and $\partial_z = 0$ then it has only one non-zero component $\vec{\omega} = \omega \vec{e}_z$ with

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{\partial}{\partial x}\left(-\frac{\partial \Psi}{\partial x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial \Psi}{\partial y}\right) = -\frac{\partial \Psi}{\partial x^2} - \frac{\partial \Psi}{\partial y^2} = -\vec{\nabla}^2 \Psi$$

The vorticity stream-function form of the Navier-Stokes equations for two- dimensional flows is the set of coupled scalar equations given below :

$$\frac{\partial \omega}{\partial t} + \vec{v} \cdot \vec{\nabla} \omega = \nu \vec{\nabla}^2 \omega \qquad \text{on} \quad \Omega \times [0, T] \tag{21.6}$$

$$\vec{\nabla}^2 \vec{\Psi} = -\omega \qquad \text{on} \quad \Omega \times [0, T] \tag{21.7}$$

where Ω is a domain of $\mathbb{R}^2$, $\vec{v}(\vec{x}, t)$ is the velocity, $\omega(\vec{x}, t)$ is the vorticity, $\Psi(\vec{x}, t)$ is the stream function, and ν is the kinematic viscosity. The vorticity and the stream function are related to the velocity field through

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{21.8}$$

$$u = \frac{\partial \Psi}{\partial y} \tag{21.9}$$

$$v = -\frac{\partial \Psi}{\partial x} \tag{21.10}$$

For flow domains having solid boundaries (e.g., walls or obstacles) the no-slip and no-penetration conditions yield constraints on the derivatives of Ψ , i.e,

$$\vec{n} \cdot \vec{\nabla} \Psi = \nu_t \tag{21.11}$$

$$\vec{t} \cdot \vec{\nabla} \Psi = 0 \tag{21.12}$$

where $\vec{t}$ and $\vec{n}$ are the tangential and normal unit vectors at the surface whereas ν_t is the tangential component of the velocity at the wall. Moreover, the stream function assumes constant values along solid boundaries. However, it is difficult to derive any boundary condition for vorticity at solid surfaces. The boundary values of the vorticity at solid surfaces are related to the local boundary-layer profiles and are thus time dependent. On the other hand, they do not arise naturally from the no-slip and no-penetration conditions.

21.2 Vorticity-stream function formulation of the isoviscous Stokes equation in 2d

In two dimensions the velocity is obtained as follows:

$$\vec{v} = (u, v) = \left(\frac{\partial \Psi}{\partial y}, -\frac{\partial \Psi}{\partial x}\right) \tag{21.13}$$

Provided the function Ψ is a smooth enough function, this automatically insures that the flow is incompressible:

$$\vec{\nabla} \cdot \vec{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \frac{\partial^2 \Psi}{\partial xy} - \frac{\partial^2 \Psi}{\partial xy} = 0 \tag{21.14}$$

Also, we have

$$\vec{v} \cdot \vec{\nabla} \Psi = \vec{v} \cdot \left(\frac{\partial \Psi}{\partial x}, \frac{\partial \Psi}{\partial y}\right) = (u, v) \cdot (-v, u) = 0$$

Assuming constant viscosity, the Stokes equation writes:

$$-\vec{\nabla}p + \eta_0 \Delta \vec{v} + \rho \vec{g} = \vec{0} \qquad (21.15)$$

or, in each dimension:

$$0 = -\frac{\partial p}{\partial x} + \eta_0 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial x^y} \right) + \rho g_x \qquad (21.16)$$

$$0 = -\frac{\partial p}{\partial y} + \eta_0 \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + \rho g_y \qquad (21.17)$$

Using Eq. (21.13), we can write

$$0 = -\frac{\partial p}{\partial x} + \eta_0 \left(\frac{\partial^3 \Psi}{\partial y^3} + \frac{\partial^3 \Psi}{\partial x^2 \partial y} \right) + \rho g_x$$

$$0 = -\frac{\partial p}{\partial y} - \eta_0 \left(\frac{\partial^3 \Psi}{\partial x^3} + \frac{\partial^3 \Psi}{\partial y^2 \partial x} \right) + \rho g_y \qquad (21.18)$$

The pressure terms in both equations can be removed by first differentiating the first line with regards to y and the second line with regards to x,

$$0 = -\frac{\partial^2 p}{\partial x \partial y} + \eta_0 \left(\frac{\partial^4 \Psi}{\partial y^4} + \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} \right) + \frac{\partial(\rho g_x)}{\partial y}$$

$$0 = -\frac{\partial^2 p}{\partial y \partial x} - \eta_0 \left(\frac{\partial^4 \Psi}{\partial x^4} + \frac{\partial^4 \Psi}{\partial y^2 \partial x^2} \right) + \frac{\partial(\rho g_y)}{\partial x} \qquad (21.19)$$

and next by subtracting the resulting equations, leading to:

$$0 = \eta_0 \left(\frac{\partial^4 \Psi}{\partial x^4} + 2\frac{\partial^4 \Psi}{\partial x^2 \partial y^2} + \frac{\partial^4 \Psi}{\partial y^4} \right) + \frac{\partial(\rho g_x)}{\partial y} - \frac{\partial(\rho g_y)}{\partial x} \qquad (21.20)$$

or

$$\vec{\nabla}^4 \Psi = \frac{1}{\eta_0} \left(-\frac{\partial(\rho g_x)}{\partial y} + \frac{\partial(\rho g_y)}{\partial x} \right) \qquad (21.21)$$

where

$$\vec{\nabla}^4 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

which we find for example in According to Gerya's book2, page 70.

Note that $\vec{\nabla}^2 \vec{\nabla}^2 = \vec{\nabla}^4$ is known as the Biharmonic operator. These equations are also to be found in the geodynamics literature, see Ismail-Zadeh and Tackley [659, eq. 1.43] or Gerya [483, p 70-71].

From $\vec{\nabla}^2 \Psi = -\omega$, we can also write

$$\vec{\nabla}^2 \omega = -\frac{1}{\eta_0} \left(-\frac{\partial(\rho g_x)}{\partial y} + \frac{\partial(\rho g_y)}{\partial x} \right) \qquad (21.22)$$

21.3 Vorticity-stream function formulation of the non-isoviscous Stokes equation in 2d

In this case, we can no longer write the Stokes equation as follows

$$-\vec{\nabla}p + \eta_0 \Delta \vec{v} + \rho \vec{g} = \vec{0} \qquad (21.23)$$

Indeed, it should be

$$-\vec{\nabla}p + \vec{\nabla}\cdot(2\eta\dot{\varepsilon}) + \rho\vec{g} = \vec{0} \tag{21.24}$$

We have

$$2\eta\dot{\varepsilon} = 2\eta \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) & \frac{\partial v}{\partial y} \end{pmatrix}$$

so that

$$W_x = -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x}\left(2\eta\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\eta(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})\right) + \rho g_x \tag{21.25}$$

$$W_y = -\frac{\partial p}{\partial y} + \frac{\partial}{\partial x}\left(\eta(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})\right) + \frac{\partial}{\partial y}\left(2\eta\frac{\partial v}{\partial y}\right) + \rho g_y \tag{21.26}$$

Using Eq. (21.13) again, we can write

$$W_x = -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x}\left(2\eta\frac{\partial^2\Psi}{\partial x\partial y}\right) + \frac{\partial}{\partial y}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \rho g_x = 0 \tag{21.27}$$

$$W_y = -\frac{\partial p}{\partial y} + \frac{\partial}{\partial x}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) - \frac{\partial}{\partial y}\left(2\eta\frac{\partial^2\Psi}{\partial y\partial x}\right) + \rho g_y = 0 \tag{21.28}$$

The pressure terms in both equations can be removed by first differentiating the first line with regards to y and the second line with regards to x,

$$\frac{\partial W_x}{\partial y} = -\frac{\partial^2 p}{\partial x\partial y} + \frac{\partial^2}{\partial x\partial y}\left(2\eta\frac{\partial^2\Psi}{\partial x\partial y}\right) + \frac{\partial^2}{\partial y^2}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \frac{\partial\rho g_x}{\partial y} \tag{21.29}$$

$$\frac{\partial W_y}{\partial x} = -\frac{\partial^2 p}{\partial y\partial x} + \frac{\partial^2}{\partial x^2}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) - \frac{\partial^2}{\partial x\partial y}\left(2\eta\frac{\partial^2\Psi}{\partial y\partial x}\right) + \frac{\partial\rho g_y}{\partial x} \tag{21.30}$$

and next by subtracting the resulting equations, leading to:

$$0 = \frac{\partial^2}{\partial x\partial y}\left(2\eta\frac{\partial^2\Psi}{\partial x\partial y}\right) + \frac{\partial^2}{\partial y^2}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) - \frac{\partial^2}{\partial x^2}\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \frac{\partial^2}{\partial x\partial y}\left(2\eta\frac{\partial^2\Psi}{\partial y\partial x}\right) + \frac{\partial\rho g_x}{\partial y} -$$

$$0 = \frac{\partial^2}{\partial x\partial y}\left(2\eta(\frac{\partial^2\Psi}{\partial x\partial y} + \frac{\partial^2\Psi}{\partial y\partial x})\right) + \left(\frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial x^2}\right)\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \frac{\partial\rho g_x}{\partial y} - \frac{\partial\rho g_y}{\partial x}$$

$$0 = 4\frac{\partial^2}{\partial x\partial y}\left(\eta\frac{\partial^2\Psi}{\partial x\partial y}\right) + \left(\frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial x^2}\right)\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \frac{\partial\rho g_x}{\partial y} - \frac{\partial\rho g_y}{\partial x} \tag{21.}$$

i.e.

$$\boxed{4\frac{\partial^2}{\partial x\partial y}\left[\eta\frac{\partial^2\Psi}{\partial x\partial y}\right] + \left(\frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial x^2}\right)\left[\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right] = -\frac{\partial\rho g_x}{\partial y} + \frac{\partial\rho g_y}{\partial x}}$$

This is eq 5.36c of Gerya. This expression (esp. the lhs) is to be found in its dimensionless form in Christensen and Yuen [263] (1984) or in Schmeling and Jacoby [1199] (1981) for example:

$$4\frac{\partial^2}{\partial x\partial y}\left(\eta\frac{\partial^2\Psi}{\partial x\partial y}\right) + \left(\frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial x^2}\right)\left(\eta(\frac{\partial^2\Psi}{\partial y^2} - \frac{\partial^2\Psi}{\partial x^2})\right) + \frac{\partial\rho g_x}{\partial y} - \frac{\partial\rho g_y}{\partial x} = Ra\frac{\partial T}{\partial x} - Rb\frac{\partial\Gamma}{\partial x}$$

In the presence of temperature variations and multiple compositions, Trim, Lowman, and Butler [1371] (2020) use the following identical nondimensional equation:

$$\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right)\left[\eta\left(\frac{\partial^2\Psi}{\partial x^2} - \frac{\partial^2\Psi}{\partial y^2}\right)\right] + 4\frac{\partial^2}{\partial xy}\left[\eta\frac{\partial^2\Psi}{\partial xy}\right] = Ra_T\frac{\partial T}{\partial x} - Ra_C\frac{\partial C}{\partial x}$$

PB: what if C is a discontinuous field? then I guess C needs to be carried on the nodes.
Introducing the vorticity again, we can rewrite the above equation as follows: ??? can it ???

21.4 Boundary conditions

When we use a velocity-pressure formulation in a 2D Cartesian domain Dirichlet boundary conditions translate into zeroing either u, v, or both, or assigning u and/or v a value on (part of) the boundary. In this case our unknowns are the Ψ function and the vorticity ω so we need to think about boundary conditions a bit more.

The solution of vorticity transport equation and stream function equation requires that appropriate vorticity and stream function boundary conditions are specified at the boundaries. The specification of these boundary conditions is extremely important since it directly affects the stability and accuracy of the solution. However, neither vorticity nor its derivatives at the boundary are usually known in advance. Therefore a set of boundary conditions must be constructed.

For example, since the flow is parallel to a solid boundary, solid boundaries and symmetry planes are surfaces of constant stream function. In other words, since flow is parallel to the walls of the cavity, walls may be treated as streamline. Thus, the stream function value on the wall streamline is set as a constant (often taken to be zero for simplicity). Since the stream function is a constant along a wall, all the derivatives of stream function along the wall also vanish.

In the end we have to account for three physical types of boundary conditions: free slip, no slip, and prescribed velocity (yes, technically the last one encompasses the second one).

Check Napolitano, Pascazio, and Quartapelle [984] (1999) for a A review of vorticity conditions in the numerical solution of the vorticity-stream function equations.

21.4.1 Free slip, 'stress free' boundary conditions

The free slip boundary condition states that at the interface between a moving fluid and a stationary wall, 1) the normal component of the fluid velocity field is equal to zero, but the tangential component is unrestricted. 2) the tangential stress is set to zero (which is why it is often called 'stress free' b.c.).

Free slip at the top and bottom then implies:

1. $v = 0$ which translates into $\frac{\partial \Psi}{\partial x} = 0$. Actually stating that $\frac{\partial \Psi}{\partial x} = 0$ along an horizontal edge means that Ψ is then constant along that edge.

2. $\tau_{xy} = 2\eta \dot{\varepsilon}_{xy} = \eta(\partial_y u + \partial_x v) = 0$ on the boundary. Since $v = 0$, then it does not change in the x direction on the boundary, i.e. $\partial_x v = 0$, leaving the condition $\partial_y u = 0$ which translates into $\frac{\partial^2 \Psi}{\partial y^2} = 0$. Since Ψ is constant on that edge (see point above) then we also have $\frac{\partial^2 \Psi}{\partial x^2} = 0$ so that we can write that $\omega = -\vec{\nabla}^2 \Psi = 0$ on that boundary.

The same reasoning on the sides means: $u = 0$ and $\partial_x v = 0$ which translates into $\frac{\partial \Psi}{\partial y} = 0$ and $\frac{\partial^2 \Psi}{\partial x^2} = 0$. In general the Poisson equation $\vec{\nabla}^2 \Psi = -\omega$ becomes on a wall:

$$\left. \frac{\partial^2 \Psi}{\partial n^2} \right|_{wall} = -\omega_{wall}$$

where n is the normal direction. Indeed, in Hsui [630] (1978) we read:

> "Besides the initial condition, boundary conditions are also required in order to form a well-posed system. Equation 13 [i.e omega = curl Psi] requires the specification of stream function at all boundaries. Since only the derivatives of stream functions which determine the velocities are important to the problem, stream function can thus be specified to within a constant. Consequently, they are chosen, for convenience, to be zero at all boundaries."

This is also coherent with for example:

$$\psi = 0 \qquad \partial^2\psi/\partial z^2 = 0 \qquad \theta = 0 \qquad \text{on} \quad z = 1$$
$$\psi = 0 \qquad \partial^2\psi/\partial z^2 = 0 \qquad \partial\theta/\partial z = 1 \qquad \text{on} \quad z = 0$$
$$\psi = 0 \qquad \partial^2\psi/\partial x^2 = 0 \qquad \partial\theta/\partial x = 0 \qquad \text{on} \quad x = 0, l$$

("free stress b.c.") Richter and McKenzie [1143]
(1983)

2.3 Boundary conditions

The boundaries of the calculation region are free-slip and stress-free; therefore boundary conditions of this situation are expressed as follows,

(1) $\phi' = 0$ at all boundaries,
(2) $\partial^2\phi'/\partial x'^2 = 0$ at both side boundaries,
(3) $\partial^2\phi'/\partial z'^2 = 0$ at top and bottom boundaries.

Matsumoto and Tomoda [890]

Fig. 1. Sketch of the model. Boundary conditions are given in dimensionless values.
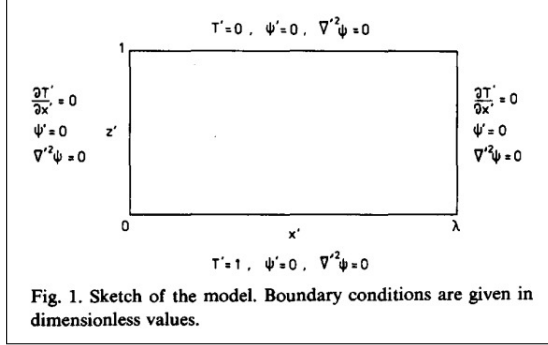
Hansen and Ebel [567]

For all experiments we used boundary conditions according to

$$z = 0, \qquad 0 < x < \lambda : \psi = 0; \qquad \nabla^2\psi = 0, \qquad T = 1.$$
$$z = 1, \qquad 0 < x < \lambda : \psi = 0; \qquad \nabla^2\psi = 0, \qquad T = 0.$$
$$0 < z < 1, \qquad x = 0 : \psi = 0; \qquad \nabla^2\psi = 0, \qquad \frac{\partial T}{\partial x} = 0. \tag{5}$$
$$0 < z < 1, \qquad x = \lambda : \psi = 0; \qquad \nabla^2\psi = 0, \qquad \frac{\partial T}{\partial x} = 0.$$

Hansen and Ebel [574]

The boundary conditions are stress-free and impermeable at the top and bottom boundaries, and reflecting along the sides. The dimensionless temperature, T, is 0 at the top and 1 at the bottom. The boundary conditions are

$$\omega(x,z) = \psi(x,z) = 0 \begin{cases} \text{for } x = 0 \text{ and } x = x_0, \\ \text{for } z = 0 \text{ and } z = 1, \end{cases} \tag{4a}$$

$$T(x,z) = \begin{cases} 1 & \text{for } z = 0, \\ 0 & \text{for } z = 1, \end{cases} \tag{4b}$$

$$\frac{\partial T}{\partial x} = 0 \text{ at } x = 0 \text{ and } x = x_0, \tag{4c}$$

where x_0 defines the aspect-ratio of the box.

Larsen, Yuen, Moser, and Fornberg [795]

21.4.2   No slip

Rather counter intuitively no-slip boundary conditions prove to be substantially more difficult to implement than free-slip boundary conditions. For example we find in DeMarco, DeAndrade, and Zaparoli [347] (2003):

> "According to Layton [800] (1999), there are at least three natural ways of imposing zero tangential velocities value along the boundary: (i) Lagrange multiplier of tangential velocity component equal to zero as a constraint; (ii) Penalty term imposing tangential velocity component equal to zero approximately; (iii) Replacing no-slip with slip with friction. "

DeMarco, DeAndrade, and Zaparoli [347] (2023) present "an analysis of penalty method application to imposing no-slip boundary condition. The approach used herein consists to express the vorticity boundary condition through the natural boundary condition depending on the solid wall tangential velocity component."

Let us consider the bottom boundary. No-slip means $u = v = 0$, i.e. $\frac{\partial\Psi}{\partial y} = \frac{\partial\Psi}{\partial x} = 0$. We have seen previously that $v = 0$ yields $\Psi = $constant on the boundary, leaving $\frac{\partial\Psi}{\partial y} = 0$. This is problematic because if one solves the biharmonic equation as two Poisson equations for ω and Ψ it does not translate to a boundary condition for ω!

In general zeroing the tangential component of the velocity on a boundary will write $\partial\Psi/\partial n = 0$ where n is the normal to the boundary.

Here u is prescribed at the top, but not zero:

Gurnis and Davies [544]

In Lux, Davies, and Thomas [863] (1979) we find this great summary of the problem and the sketch of its solution:

"The numerical methods we have used are well established and easy to apply. If we introduce the vorticity ω, then the fourth-order stream function equation can be separated into a pair of second-order Poisson equations,

$$\omega = -\vec{\nabla}^2\Psi \qquad \vec{\nabla}^2\omega = -\frac{\partial T}{\partial x},$$

for which efficient matrix reduction methods are available. For prescribed stress boundary conditions, the Poisson equations can simply be solved successively as done by McKenzie, Roberts, and Weiss [906] (1974). A complication arises for a prescribed boundary velocity because the boundary conditions on vorticity are then not explicit, see Richter [1142] (1973). On the boundary the vorticity is proportional to the shear stress and, since the shear stress is not known until the flow field is found, we lack a boundary condition on ω. The Poisson equation for the stream function has an over-prescribed boundary condition, since both Ψ, and $\partial\Psi/\partial y$ are given. This dilemma is resolved by using the condition on Ψ for the stream function equation and the condition on $\partial\Psi/\partial y$ in a Taylor series expansion for ω at the boundary for the vorticity equation. To ensure numerical stability with this boundary condition, it is necessary to iterate between the equations; an efficient way of optimizing this iteration has been given by Ehrlich and Gupta [386] (1975). The iteration technique for the coupled Poisson equations was tested against analytical solutions of the biharmonic equation given by Davies [333] (1977) and shown to be very accurate. The solution of the convection problem described here evolves with time, so the boundary condition on ω must be found at each time step by the iteration method. This procedure is time consuming, so to increase the rate of convergence of this iteration a boundary condition on ω accurate to first order (Taylor series) is used."

In Comini, Manzan, and Nonino [294] (1994) we read:

In the analysis of two-dimensional incompressible flows the streamfunction-vorticity formulation of the Navier-Stokes equations allows the elimination of pressure from the problem and automatically satisfies the continuity constraint. On the other hand, the value of the vorticity at no-slip boundaries is difficult to specify and a poor evaluation of this boundary condition leads, almost invariably, to serious difficulties in obtaining a converged solution. [...] A guideline for a correct specification of boundary conditions at no-slip walls has been given by Roache.' His numerical recipe can be summarized as follows: at no-slip wallsjirst specify the streamfunction and then, in the procurement of the wall vorticity, utilize the additional information on the normal component of the streamfunction. In this way the boundary conditions for the streamfunction are not overspecified and the wall vorticity is correctly related to the tangential component of the velocity. Obviously, at stationary walls the tangential component of the velocity is zero, but at moving walls serious errors may result if the gradient of the streamfunction is not properly taken into account.

21.4.3 Stress b.c.

To Do

21.4.4 Line of symmetry

When the flow is truly symmetrical, the axis of symmetry can be considered a streamline. Therefore the value of streamfunction along this boundary can be specified. Obviously, the velocity component normal to the the symmetry boundary would be zero, whereas the streamwise component is extrapolated from the interior solution.

21.5 Pressure Poisson Equation for the isoviscous N-S equation

This section is inspired by Salih [1172].

The PPE can be derived by taking the divergence of vector form of momentum equation

$$\vec{\nabla} \cdot \left[\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} \right] = \vec{\nabla} \cdot \left[-\frac{1}{\rho}\vec{\nabla}p + \nu\vec{\nabla}^2\vec{v} \right] \tag{21.32}$$

or,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{21.33}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial y} + \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) \tag{21.34}$$

Differentiating these equations with respect to x and y respectively, and assuming ρ constant in space yields

$$\frac{\partial^2 u}{\partial x\partial t} + \frac{\partial u}{\partial x}\frac{\partial u}{\partial x} + u\frac{\partial^2 u}{\partial x^2} + \frac{\partial v}{\partial x}\frac{\partial u}{\partial y} + v\frac{\partial^2 u}{\partial x\partial y} = -\frac{1}{\rho}\frac{\partial^2 p}{\partial x^2} + \nu\frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{21.35}$$

$$\frac{\partial^2 v}{\partial y\partial t} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + u\frac{\partial^2 v}{\partial y\partial x} + \frac{\partial v}{\partial y}\frac{\partial v}{\partial y} + v\frac{\partial^2 v}{\partial y^2} = -\frac{1}{\rho}\frac{\partial^2 p}{\partial y^2} + \nu\frac{\partial}{\partial y}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) \tag{21.36}$$

Adding both equations together, i.e. taking the divergence of the first equation, yields

$$\frac{\partial^2 u}{\partial x \partial t} + \frac{\partial u}{\partial x}\frac{\partial u}{\partial x} + u\frac{\partial^2 u}{\partial x^2} + \frac{\partial v}{\partial x}\frac{\partial u}{\partial y} + v\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial y \partial t} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + u\frac{\partial^2 v}{\partial y \partial x} + \frac{\partial v}{\partial y}\frac{\partial v}{\partial y} + v\frac{\partial^2 v}{\partial y^2}$$

$$= -\frac{1}{\rho}\frac{\partial^2 p}{\partial x^2} + \nu\frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) - \frac{1}{\rho}\frac{\partial^2 p}{\partial y^2} + \nu\frac{\partial}{\partial y}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) \qquad (21.37)$$

$$\partial_t(\underbrace{\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}}_{=0}) + \frac{\partial u}{\partial x}\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + \frac{\partial v}{\partial x}\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}\frac{\partial v}{\partial y} + u(\underbrace{\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y \partial x}}_{A}) + v(\underbrace{\frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 u}{\partial x \partial y}}_{B})$$

$$= -\frac{1}{\rho}(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}) + \nu\frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \nu\frac{\partial}{\partial y}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) \qquad (21.38)$$

We have

$$A = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y \partial x} = \frac{\partial u}{\partial x}(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}) = 0$$

$$B = \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 u}{\partial x \partial y} = \frac{\partial v}{\partial y}(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}) = 0$$

so that

$$(\frac{\partial u}{\partial x})^2 + 2\frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + (\frac{\partial v}{\partial y})^2 = -\frac{1}{\rho}(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}) + \nu\frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \nu\frac{\partial}{\partial y}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) \quad (21.39)$$

The viscosity terms in the rhs can be rearranged as

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \frac{\partial}{\partial y}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) = \frac{\partial^2}{\partial x^2}(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}) + \frac{\partial^2}{\partial y^2}(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}) = 0$$

We are left with

$$(\frac{\partial u}{\partial x})^2 + 2\frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + (\frac{\partial v}{\partial y})^2 = -\frac{1}{\rho}(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}) \qquad (21.40)$$

Now, the left-hand side can be further reduced as follows

$$(\frac{\partial u}{\partial x})^2 + 2\frac{\partial u}{\partial y}\frac{\partial v}{\partial x} + (\frac{\partial v}{\partial y})^2 = (\underbrace{\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}}_{=0})^2 - 2\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + 2\frac{\partial u}{\partial y}\frac{\partial v}{\partial x}$$

In the end

$$\vec{\nabla}^2 p = 2\rho\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}\frac{\partial u}{\partial y}\right)$$

The PPE can also be written in terns of stream function:

$$\boxed{\vec{\nabla}^2 p = 2\rho\left[\frac{\partial^2 \Psi}{\partial x^2}\frac{\partial^2 \Psi}{\partial y^2} - \left(\frac{\partial^2 \Psi}{\partial x \partial y}\right)^2\right]}$$

The Poisson equation for pressure is an elliptic equation, showing the elliptic nature of pressure in incompressible flows. For a steady flow problem, the PPE is solved only once, i.e., after the steady state values of ω and Ψ have been computed.

Having established the PDE we must now turn to the boundary conditions. On a solid boundary, boundary values of pressure are obtained using the tangential momentum equation of the fluid

adjacent to the wall surface. For a wall located at $y = 0$ in Cartesian coordinate system (i.e. the bottom boundary), the tangential momentum equation, that is the x-momentum equation, reduces to

$$\left.\frac{\partial p}{\partial x}\right|_{wall} = \eta \left.\frac{\partial^2 u}{\partial y^2}\right|_{wall} = -\eta \left.\frac{\partial \omega}{\partial y}\right|_{wall}$$

since $v = \partial_x v = 0$ on that boundary.

In Reddy and Gartling [1118] we find at page 163: The pressure, if required, may be computed from a Poisson equation of the form:

$$\Delta P = \vec{\nabla} \cdot (\vec{f} - \nu \Delta \vec{v} - (\vec{v} \cdot \vec{\nabla})\vec{v})$$

Check Gunzburger and Peterson [538] (1988) which "describe an alternate method of recovering the pressure which does not encounter any of the above difficulties; in particular, absolutely no boundary conditions on the pressure are needed at solid walls."

21.6 Vorticity-stream function formulation in 3D

The vorticity-streamfunction approach has seen considerable use for two-dimensional incompressible flows. It has become less popular in recent years because its extension to three-dimensional flows is difficult. Both the vorticity and streamfunction become three-component vectors in three dimensions so one has a system of six partial differential equations in place of the four that are necessary in a velocity-pressure formulation. It also inherits the difficulties in dealing with variable fluid properties, compressibility, and boundary conditions that were described above for two dimensional flows.

In Reddy and Gartling [1118] we find at page 163:

"A 3D version of the vorticiy transport equation is possible, although it has seen relatively little use in computation due to the complexity of the vorticity boundary conditions."

Glaisner and Tezduyar [493] (1987) state: "

The extension of the codes to three dimensions is not easy for the vorticity stream-function formulation. In three dimensions, the vorticity- stretching term does not vanish and the equation system has two nonlinear terms."

In Zhong, Yuen, Moresi, and Knepley [1524] (2012) we read:

"For 3-D mantle convection, we can likewise employ the poloidal potential Φ and a vorticity-like scalar function Ω (Busse, 1989; Chandrasekhar, 1961; Travis et al., 1990). It is to be noted that this potential Φ is not the same as the stream function Ψ in 2-D (see Chapter 7.04). From the general representation of an arbitrary solenoidal vector field (Busse, 1989), we can write a 3-D velocity vector as

$$\vec{v} = \vec{\nabla} \times \vec{\nabla} \times (\Phi \vec{e}_z) + \vec{\nabla} \times (\Theta \vec{e}_z)$$

where $\vec{e}_z$ is the unit vector in the vertical, z-direction, pointing upward. Θ is the toroidal potential and is present in problems with lateral variations of viscosity (Christensen and Harder, 1991; Gable et al., 1991; Zhang and Christensen, 1993). Thus, for constant viscosity, Θ is zero unless driven by a boundary condition (e.g., Gable et al., 1991), and the velocity vector $\vec{v} = (u, v, w)$ involves higher-order derivatives of Φ in this formulation:

$$u = \frac{\partial^2 \Phi}{\partial y \partial z} \qquad v = \frac{\partial^2 \Phi}{\partial x \partial z} \qquad w = -\left(\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2}\right)$$

The 3-D momentum equation for constant properties can be written as a system of coupled Poisson equations in 3-D:

$$\vec{\nabla}^2 \Phi = \Omega \tag{21.41}$$

$$\vec{\nabla}^2 \Omega = \mathsf{Ra}T \tag{21.42}$$

where all differential operators are 3-D in character and Ω is a scalar function playing a role analogous to the vorticity in the 2-D formulation.

We note that FD and FV methods using the primitive variables of velocity and pressure are currently predominant in models of 3-D mantle convection with variable viscosity. ”

In Rabinowicz, Rouzo, Sempere, and Rosemberg [1098] (1993) we read:

The velocity field being solenoidal, it may be written as the sum of a poloidal and a toroidal field derived from the scalar functions Q and S such that [Chandrasekhar,1961]

$$u = \frac{\partial^2 Q}{\partial x \partial z} + \frac{\partial S}{\partial y} \tag{21.43}$$

$$v = \frac{\partial^2 Q}{\partial y \partial z} - \frac{\partial S}{\partial x} \tag{21.44}$$

$$w = -\left(\frac{\partial^2 Q}{\partial x^2} + \frac{\partial^2 Q}{\partial y^2}\right) \tag{21.45}$$

After elimination of the pressureterms by taking the curl of equation (1), we obtain

$$\mu_i \vec{\nabla}^4 Q = -\mathsf{Ra}\theta \tag{21.46}$$

$$\vec{\nabla}^2 S = 0 \tag{21.47}$$

$$p = \mu_i \vec{\nabla}^2 \frac{\partial Q}{\partial z} \tag{21.48}$$

More details about the notations and the boundary conditions are provided in the paper.

ToDo:

1. scan 3D literature

2. find all formulations cartesian or not

3. document num methods used

21.6.1   Cartesian domain (1)

Sotin and Labrosse [1260] (1999) write:

“ The conservation equations for momentum and mass are transformed into four Poisson equations by introducing a vector potential stream function Ψ and vorticity ω [1370]:

$$\Delta \Psi_x + \omega_x = 0 \tag{21.49}$$

$$\Delta \omega_x - \mathsf{Ra}_T \frac{\partial T}{\partial y} = 0 \tag{21.50}$$

$$\Delta \Psi_y + \omega_y = 0 \tag{21.51}$$

$$\Delta \omega_y + \mathsf{Ra}_T \frac{\partial T}{\partial x} = 0 \tag{21.52}$$

Since the buoyancy force acts only in the z-direction, Ψ_z and ω_z vanish identically. Shear-stress free boundaries on top and bottom yields the following boundary conditions:

$$\omega_x = \omega_y = \Psi_x = \Psi_y = 0$$

at the top and bottom. Periodicity is assumed on all vertical boundaries.

The four Poisson equations are solved using a multigrid iterative method described by Parmentier, Sotin, and Travis [1036] (1994) and Sotin et al. 1995. "

21.6.2 Cartesian domain (2)

Let us now turn to Larsen, Yuen, Malevsky, and Smedsmo [793] (1996):

"For the spatial discretization fourth-order correct bi-cubic splines are employed [873]. [...] The mesh is uniform in the horizontal direction, and non-uniform in the vertical direction with mesh-refinement near the boundary layers. [...] The mechanical boundary conditions are stress-free and impermeable at the top and bottom boundaries, and reflecting along the sides. The dimensionless temperature T is zero at the top and unity at the bottom ($z = 1$), and there is zero heat flux along the sides."

In Larsen, Yuen, Moser, and Fornberg [795] (1997) the authors start with a 2D formulation:

"Instead of the biharmonic equation for Ψ, (e.g., Christensen [261], 1984), the conservation equations for the mass and momentum are given by two second-order partial differential equations (1) and (2). The horizontal and vertical coordinates are, respectively, x and z with z pointing upwards. Time, t, has been non-dimensionalized by the thermal diffusion time across the depth of the layer. A more detailed description of the scheme for non-dimensionalization is given in Weinstein, Olson, and Yuen [1447] (1989). The boundary conditions are stress-free and impermeable at the top and bottom boundaries, and reflecting along the sides. The dimension- less temperature, T, is 0 at the top and 1 at the bottom. The boundary conditions are: $\Psi = \omega = 0$ on all sides."

They go further and explain the 3D approach:

" We employ the poloidal potential, Ψ, and a vorticity-like scalar function Ω for solving the momentum equation with constant properties (Busse [208], 1989; Travis, Olson, and Schubert [1370], 1990b). We note that the potential Ψ is not the same as the stream-function in two dimensions. From the general representation of an arbitrary solenoidal vector field (Busse [208], 1989), we can write a three-dimensional velocity vector,

$$\vec{v} = \vec{\nabla} \times (\vec{\nabla} \times \vec{e}_z \Psi)$$

with the $\vec{e}_z$ axis pointing upwards. Thus the velocity vector involves higher-order derivatives of in this formulation:

$$u = \frac{\partial^2 \Psi}{\partial x \partial z} \qquad v = \frac{\partial^2 \Psi}{\partial y \partial z} \qquad \omega = -\left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} \right)$$

with $\vec{v} = (u, v, w)$. The three-dimensional momentum equation can be written as a system of coupled Poisson equations:

$$\vec{\nabla}^2 \Psi = \Omega \qquad \vec{\nabla}^2 \Omega = \mathsf{Ra} T$$

Here Ω is a scalar function playing an analogous role to the vorticity in the two-dimensional formulation.

The numerical method employed in this paper is based on a recursive algorithm for generating the weights in the finite-difference approximation."

21.6.3 Cartesian domain (3)

See for example Houseman [626] (1990) and refs therein:

"As the flow is incompressible, velocity may be expressed as the curl of a solenoidal vector potential $\vec{\Psi}$ [e.g. Richardson & Cornish (1977)]:

$$\vec{v} = \vec{\nabla} \times \vec{\Psi}$$

For 2-D flow, 3 is a vector everywhere normal to the plane of the flow. It may then be treated as a scalar, and is given the name streamfunction.

The momentum equation may be restated as a biharmonic equation:

$$\vec{\nabla}^4 \vec{\Psi} = \vec{f}$$

where $\vec{f}$ is minus the curl of the buoyancy force divided by viscosity, and is hereafter treated as a known quantity. It is also useful to define the vorticity, which is simply related to the potential function

$$\vec{\omega} = \vec{\nabla} \times \vec{v} = -\vec{\nabla}^2 \Psi$$

We now consider a finite difference technique for solving equation (3) in a region spanned by a regular isotropic array of nodes."

Boundary conditions of all kinds are extensively discussed!

21.6.4 Spherical shell

Zebib, Schubert, and Straus [1503] (1980) present axisymmetric steady convective solutions
    Likewise Solheim and Peltier [1258] and Solheim and Peltier [1256]: "The model is spherical but restricted in generality to the analysis of axisymmetric solutions."

21.7 Vorticity-stream function formulation in polar/cylindrical coordinates

`https://en.wikipedia.org/wiki/Stokes_stream_function` uses very different definitions! sort it out! actually it is for three-dimensional incompressible flow with axisymmetry!

$$v_r = \frac{1}{r}\frac{\partial \Psi}{\partial \theta}$$

$$v_\theta = -\frac{\partial \Psi}{\partial r}$$

$$\vec{\omega} = \left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\vec{e}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\vec{e}_\theta + \left(\frac{1}{r}\frac{\partial(rv_\theta)}{\partial r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)\vec{e}_z$$

$$\vec{\nabla}^4 \Psi = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right)\left(\frac{\partial^2 \Psi}{\partial r^2} + \frac{1}{r}\frac{\partial \Psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \Psi}{\partial \theta^2}\right)$$

or

$$\vec{\nabla}^4 \Psi = \Psi_{,rrrr} + \frac{2}{r}\Psi_{,rrr} - \frac{1}{r^2}(\Psi_{,rr} - 2\Psi_{,rr\theta\theta}) + \frac{1}{r^3}(\Psi_{,r} - 2\Psi_{,r\theta\theta}) + \frac{1}{r^4}(4\Psi_{,\theta\theta} + 2\Psi_{,\theta\theta\theta\theta})$$

See Hsui [630]. [1367] [1369]

21.8 Vorticity-stream function formulation in spherical coordinates for three-dimensional incompressible flow with axisymmetry

What follows comes from `https://en.wikipedia.org/wiki/Stokes_stream_function`.

The flow velocity components $\mathbf{v}_r$ and $\mathbf{v}_\theta$ are related to the Stokes stream function Ψ through:

$$\mathbf{v}_r = \frac{1}{r^2 \sin\theta} \frac{\partial \Psi}{\partial \theta} \qquad \mathbf{v}_\theta = -\frac{1}{r \sin\theta} \frac{\partial \Psi}{\partial r}$$

The azimuthal velocity component $\mathbf{v}_\phi$ is not a function of the Stokes stream function Ψ.
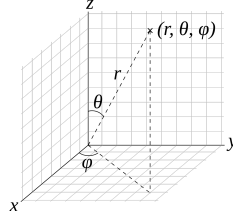
The vorticity is defined as:

$$\vec{\omega} = \vec{\nabla} \times \vec{\mathbf{v}} = \vec{\nabla} \times \vec{\nabla} \times \Psi \qquad \text{with} \qquad \Psi = -\frac{\Psi}{r \sin\theta} \vec{e}_\phi$$

From the definition of the curl in spherical coordinates[2]:

$$\omega_r = \frac{1}{r \sin\theta} \left(\frac{\partial}{\partial \theta} (\mathbf{v}_\phi \sin\theta) - \frac{\partial \mathbf{v}_\theta}{\partial \phi} \right) \vec{e}_r \tag{21.53}$$

$$\omega_\theta = \frac{1}{r} \left(\frac{1}{\sin\theta} \frac{\partial \mathbf{v}_r}{\partial \phi} - \frac{\partial}{\partial r} (r \mathbf{v}_\phi) \right) \vec{e}_\theta \tag{21.54}$$

$$\omega_\phi = \frac{1}{r} \left(\frac{\partial}{\partial r} (r \mathbf{v}_\theta) - \frac{\partial \mathbf{v}_r}{\partial \theta} \right) \vec{e}_\phi \tag{21.55}$$

First notice that the r and θ components are equal to 0. prove! Secondly substitute $\mathbf{v}_r$ and $\mathbf{v}_\theta$ into ω_ϕ. The result is:

$$\omega_r = 0 \tag{21.56}$$
$$\omega_\theta = 0 \tag{21.57}$$
$$\omega_\phi = \frac{1}{r} \left[\frac{\partial}{\partial r} \left(r \left(-\frac{1}{r \sin\theta} \frac{\partial \Psi}{\partial r} \right) \right) - \frac{\partial}{\partial \theta} \left(\frac{1}{r^2 \sin\theta} \frac{\partial \Psi}{\partial \theta} \right) \right] \tag{21.58}$$

After some algebra we arrive at

$$\vec{\omega} = \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{r \sin\theta} \left(\frac{\partial^2 \Psi}{\partial r^2} + \frac{\sin\theta}{r^2} \frac{\partial}{\partial \theta} \left(\frac{1}{\sin\theta} \frac{\partial \Psi}{\partial \theta} \right) \right) \end{pmatrix}$$

Proof that velocity is perpendicular to gradient of stream function:

$$\vec{\nabla}\Psi \cdot \vec{\mathbf{v}} = \frac{\partial \Psi}{\partial r} \cdot \frac{1}{r^2 \sin\theta} \frac{\partial \Psi}{\partial \theta} + \frac{1}{r} \frac{\partial \Psi}{\partial \theta} \cdot \left(-\frac{1}{r \sin\theta} \frac{\partial \Psi}{\partial r} \right) = 0$$

---

[2]`https://en.wikipedia.org/wiki/Del_in_cylindrical_and_spherical_coordinates`

21.9 Numerical approach

When it comes to solving the biharmonic equation, there are essentially two options: solving/discretising the biharmonic operator involving 4th order derivatives, or introducing the vorticity and solving coupled PDEs in vorticity-stream function.

Malevsky and Yuen [875] state:

> " The fourth-order elliptic equation does not contain time explicitly. One can obtain the stream-function for a given temperature field by solving this nonlinear equation at each instant. Equation (12) is a nonlinear time-dependent advection- diffusion equation, where the $\Psi(T)$ dependence is given by the biharmonic eq. Numerical solution of the advection-diffusion equation with advection dominating over diffusion is fraught with numerical difficulties. It is known that high-order finite-element and finite- difference schemes deteriorate from errors located near steep gradients of the advected field. Various upwinding schemes, designed to overcome this difficulty, introduce artificial smoothing (Cuvelier et al., 1988). A finite-element scheme based on the Lagrangian formulation of the total time derivative in the advection-diffusion equation [873] proves to be very efficient for the high Rayleigh number Newtonian convection. In stress-softening fluid, thermal advection can be very strong locally due to the decrease of the effective viscosity with growing stress. Therefore, we have chosen the Lagrangian scheme to solve the energy equation for the non-Newtonian case. A second-order predictor-corrector scheme with Lagrangian formulation was applied for time-stepping in the energy equation. The Lagrangian scheme requires interpolation to compute values of Ψ and T between nodal points. This scheme is very sensitive to the quality of spatial approximation. We used bicubic splines (Ahlberg et al., 1967) for spatial discretization of both the temperature and stream-function. "

Rather interestingly Poliakov and Podlachikov [1073] (1992) state:

> " There is a class of methods based on the introduction of a stream function which satisfies this condition automatically and gives good results. However, the stress-free boundary condition causes significant difficulties because it requires the calculation of third-order derivatives of the stream function. These algorithms are also not applicable for irregular meshes. Thus, it is difficult to solve problems in regions with complicated geometry, and to refine the mesh in areas of special interest. A further disadvantage is that the approximation functions are not expressed explicitly in terms of the nodal variables and an additional system of linear equations must be solved. "

At the moment (and before a thorough literature review) here is how I see things when it comes to numerical methods:

```
                    Biharmonic equation
                             |
         ------------------------------------------
             |                                   |
         -------------                       ------------
          split into                          solve as is
         2 Poisson eqs                       ------------
         -------------                        |         |
          |         |                         FE        FD
         FE        FD          |---------|              |
                            B-splines    regular
```

1141
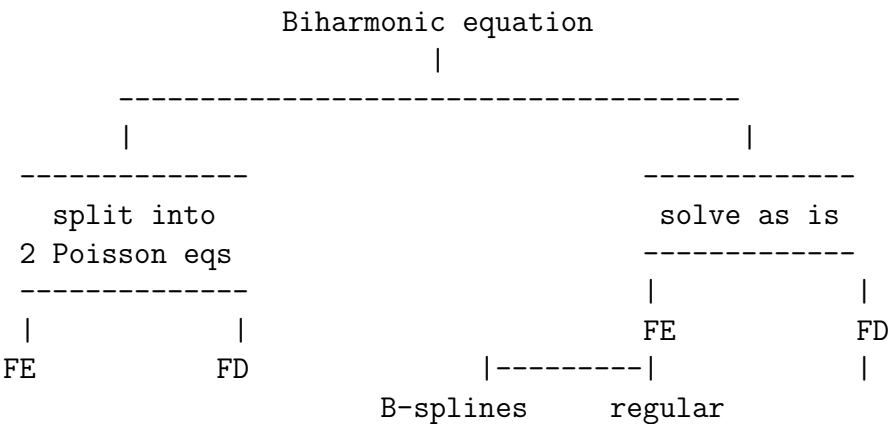
21.9.1 Finite differences

In Valera, Negredo, and Villaseñor [1392] we find

> "To solve the equation of motion we have applied a second-order, central finite- difference scheme. We considered free slip (zero shear stress) boundary conditions at all boundaries. The solution of the equation on each node was then computed in terms of the 12 nearest nodes. The system matrix was square, symmetric and diagonally dominant, with only 12 non-null diagonals. This system could be solved by a very robust method based on a LU triangular factorization by Gaussian elimination with partial pivoting, which is already implemented in the standard MATLAB code."

In Zhong, Yuen, Moresi, and Knepley [1524] (2012) we read:

> "We note that solving the biharmonic equation by FDs takes more time than solving two coupled Laplacian equations."

ToDo:

1. scan literature. e.g. McKenzie, Roberts, and Weiss [906] (1974) explain their stencils

2. establish stencil(s) for equations

3. how to deal with boundaries

4. In gerya's book it is solved with FD. see page 72, and example Streamfunction2D.m - translate to python?

21.9.2 Finite elements

We start from $\vec{\nabla}^2\vec{\nabla}^2\Psi = f$, which we rewrite as follows:

$$\vec{\nabla}^2\Psi = -\omega \tag{21.59}$$
$$\vec{\nabla}^2\omega = f \tag{21.60}$$

After establishing the weak form and discretising it, this will yield the following linear system:

$$\begin{pmatrix} K & M \\ 0 & K \end{pmatrix} \cdot \begin{pmatrix} \vec{\Psi} \\ \vec{\omega} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{f} \end{pmatrix}$$

where $K = \int B^T B dV$. This is an ideal situation: one can first solve the second line, obtain $\vec{\omega}$ and then solve the first line as $K \cdot \vec{\Psi} = -M \cdot \vec{\omega}$ Also it is the same matrix K, only different rhs!
Can i do this with Q1 elements only?!
Remarks:

- See Gresho and Sani [517] p523-525 for important remarks about establishing the weak forms of the vorticity-stream function equations. They refer to [1296].

- In Step 47 of deal.II[3] it is explained

---

[3] https://www.dealii.org/current/doxygen/deal.II/step_47.html

"The fundamental issue with the equation is that [the biharmonic equation] takes four derivatives of the solution. In the case of the Laplace equation [...], and several other tutorial programs, one multiplies by a test function, integrates, integrates by parts, and ends up with only one derivative on both the test function and trial function - something one can do with functions that are continuous globally, but may have kinks at the interfaces between cells: The derivative may not be defined at the interfaces, but that is on a lower-dimensional manifold (and so doesn't show up in the integrated value).

But for the biharmonic equation, if one followed the same procedure using integrals over the entire domain (i.e., the union of all cells), one would end up with two derivatives on the test functions and trial functions each. If one were to use the usual piecewise polynomial functions with their kinks on cell interfaces, the first derivative would yield a discontinuous gradient, and the second derivative with delta functions on the interfaces - but because both the second derivatives of the test functions and of the trial functions yield a delta function, we would try to integrate the product of two delta functions."

Logically the biharmonic equation is then split into two Poisson equations.

- In van Keken's phd thesis we read:

  "The equation of motion is a 4th-order PDE and, for a FE approximation to be conform, the basis functions should be twice continuously differentiable in each element and least continuously differentiable throughout the computational domain Ω. "

van Keken then proceeds to use 2 different methods: a non-conforming type of element Hansen and Ebel [567] (1984), and a bicubic spline FE method Woidt [1466] (1978), Kopitzke et al. [766] (1979).

- check Comini, Manzan, and Nonino [294] (1994) for N-S implementation.

- check Gunzburger and Peterson [538] (1988) for 2D, 3D implementation

the Spline FE method

Christensen [261] (1984) writes:

"The numerical grid consists of rectangular elements of variable size. All spatially varying quantities - i.e. Ψ, T and η are represented by polynomial splines. These are polynomials of a certain degree n within each individual element which are continuous at the joints up to the kth derivative ($k < n$). In a spline space at a rather compact base can be constructed, which means that each base function (B spline) is zero outside a small region consisting of few elements; this is an advantage over a polynomial or Fourier space. The shape of each B spline depends on the local grid structure. With the restriction that the size of the elements varies in both dimensions only by one of the factors $1/\alpha$, 1 or α compared to the neighbouring elements (with α fixed) a restricted number of types of B splines are necessary which reduces the computational costs. Splines which are non-zero at the boundary of the mathematical domain are modified in such a way that the boundary conditions to [the biharmonic equation] are fulfilled for any function that can be constructed in the spline space. Compared to the normal finite element method with Lagrangian shape functions, higher requirements on the continuity can be easily fulfilled. On the other hand one is restricted to semi-regular grids with all element boundaries parallel or perpendicular to each other. [...] The integral contains fourth-order

1143

derivatives, however, by partial integration it can be transformed into one containing only second derivatives. The matrix built is band structured and a stable solution is obtained by a Cholesky transformation."

Christensen and Yuen [264] (1989) states:

"The spline finite element method used to solve the set of equations (7) and (9) is described by Christensen [1984]. Bicubic splines are taken for Ψ,and biquadratic splines for T. The grid is nonuniform and allows higher resolution in the boundary layers"

In Malevsky and Yuen [875] we find:

"We can use the Galerkin method to approximate the solution. In this case a matrix $\boldsymbol{A}$ is obtained as

$$a_{ij} = \iint \eta \left[\left(\frac{\partial^2 B_i}{\partial y^2} - \frac{\partial^2 B_i}{\partial x^2} \right) \left(\frac{\partial^2 B_j}{\partial y^2} - \frac{\partial^2 B_j}{\partial x^2} \right) + 4 \frac{\partial B_i}{\partial y \partial x} \frac{\partial B_j}{\partial y \partial x} \right] dx dy$$

where B is a basic bicubic spline. The derivatives of a basic spline (a piecewise cubic polynomial) can be calculated analytically."

ToDo:

- derive B spline equation, implement, python fct? check for example [766]

- derive equation above

- work out the 2 formulations of vKK phd thesis

- check book by Gunzburger & Peterson for splines + FEM

21.10 Algorithm for stream function-vorticity formulation

A solution algorithm for computing evolution of incompressible, two-dimensional flow using stream function-vorticity formulation is given as follows:

1. Initialize the velocity field and compute the associated vorticity field and streamfunction field using equations $\omega_z = \partial_x v - \partial_y u$ and $\Delta \Psi = -\omega$.

2. Compute the boundary conditions for vorticity.

3. Solve the vorticity transport equation (21.6) to compute the vorticity at new time step; any standard time marching scheme may be used for this purpose.

4. Solve the Poisson equation for streamfunction $\Delta \Psi = -\omega$ to compute the streamfunction field at new time step; any iterative scheme for elliptic equations may be used.

5. Compute the velocity field at new time step using the relations $u = \partial_y \Psi$ and $v = -\partial_x \Psi$.

6. Return to step 2 and repeat the computation for another time step.

21.11 The nondimensional equations

21.11.1 Isoviscous case

We start from the equations (3.81),(3.82),(3.83):

$$-\vec{\nabla}p + \vec{\nabla} \cdot 2\dot{\varepsilon} + \mathsf{Ra}T\vec{e}_y = \vec{0} \tag{21.61}$$

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{21.62}$$

$$\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla}T = \kappa \Delta T \tag{21.63}$$

If the viscosity is constant then $\eta = \eta_0 = \eta_{ref}$ so $\eta = 1$ we can also write these as

$$-\vec{\nabla}p + \vec{\nabla}^2\vec{v} + \mathsf{Ra}T\vec{e}_y = \vec{0} \tag{21.64}$$

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{21.65}$$

$$\frac{\partial T}{\partial t} + \vec{v} \cdot \vec{\nabla}T = \kappa \Delta T \tag{21.66}$$

Following the approach in Section 21.2, and omitting the colors to indicate dimensionless values: or, in each dimension:

$$0 = -\frac{\partial p}{\partial x} + \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial x^y}\right) \tag{21.67}$$

$$0 = -\frac{\partial p}{\partial y} + \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) + \mathsf{Ra}T \tag{21.68}$$

Using Eq. (21.13), we can write

$$0 = -\frac{\partial p}{\partial x} + \left(\frac{\partial^3 \Psi}{\partial y^3} + \frac{\partial^3 \Psi}{\partial x^2 \partial y}\right)$$

$$0 = -\frac{\partial p}{\partial y} - \left(\frac{\partial^3 \Psi}{\partial x^3} + \frac{\partial^3 \Psi}{\partial y^2 \partial x}\right) + \mathsf{Ra}T \tag{21.69}$$

The pressure terms in both equations can be removed by first differentiating the first line with regards to y and the second line with regards to x,

$$0 = -\frac{\partial^2 p}{\partial x \partial y} + \left(\frac{\partial^4 \Psi}{\partial y^4} + \frac{\partial^4 \Psi}{\partial x^2 \partial y^2}\right)$$

$$0 = -\frac{\partial^2 p}{\partial y \partial x} - \left(\frac{\partial^4 \Psi}{\partial x^4} + \frac{\partial^4 \Psi}{\partial y^2 \partial x^2}\right) + \frac{\partial(\mathsf{Ra}T)}{\partial x} \tag{21.70}$$

and next by subtracting the resulting equations, leading to:

$$0 = \left(\frac{\partial^4 \Psi}{\partial x^4} + 2\frac{\partial^4 \Psi}{\partial x^2 \partial y^2} + \frac{\partial^4 \Psi}{\partial y^4}\right) - \mathsf{Ra}\frac{\partial T}{\partial x} \tag{21.71}$$

or

$$\vec{\nabla}^4 \Psi = \mathsf{Ra}\frac{\partial T}{\partial x} \tag{21.72}$$

Introducing again the vorticity $\omega = -\vec{\nabla}^2\Psi$, then we must solve

$$\vec{\nabla}^2\omega = -\mathsf{Ra}\frac{\partial T}{\partial x} \qquad \text{and} \qquad \vec{\nabla}^2\Psi = -\omega$$

---

check further

The vorticity equation is

$$\frac{\partial \omega}{\partial t} + \vec{\mathsf{v}} \cdot \vec{\nabla}\omega = Re^{-1}\vec{\nabla}^2\omega$$

ToDo:

- re-derive these equations , see for example Solheim and Peltier [1256] (1994) eq 8, or Kopitzke et al. [766] eq 1,2

21.12   Incorporation of phase changes

ToDo:

- scan literature for Γ fct

- isolate example

21.13   The energy equation

The (simple) energy equation

$$\frac{\partial T}{\partial t} + \vec{\mathsf{v}} \cdot \vec{\nabla}T = \kappa\Delta T$$

becomes in 2D:

$$\frac{\partial T}{\partial t} + \frac{\partial \Psi}{\partial y}\frac{\partial T}{\partial x} - \frac{\partial \Psi}{\partial x}\frac{\partial T}{\partial y} = \kappa\Delta T$$

rewrite with all coeffs

21.14   Remark, misc

Arie van den Berg (prov. comm.) writes:

> "The stream function formulation is useful to obtain a conceptually simpler implementation with finite differences if you rewrite the 4th order biharmonic equation for the streamfunction in two coupled Poisson equations for streamfunction and vorticity respectively. That is for an isoviscous fluid. For variable viscosity it becomes more ugly.
>
> It disappeared probably due its limitations for variable viscosity fluids and once the more power full but complex formulation alternatives became more well known. "

In Gresho and Sani [517] (p. 941), the authors present an algorithm to compute Ψ from a velocity field obtained via the FEM.

21.15   Literature

1961 •Hsaio-Lan Kuo. "Solution of the non-linear equations of cellular convection and heat transport". In: *Journal of Fluid Mechanics* 10.4 (1961), pp. 611–634. DOI: 10.1017/S0022112061000408

1967 •DL Turcotte and ER Oxburgh. "Finite amplitude convective cells and continental drift". In: *Journal of Fluid Mechanics* 28.1 (1967), pp. 29–42. DOI: 10.1017/S0022112067001880

1969 •G Schubert, DL Turcotte, and ER Oxburgh. "Stability of planetary interiors". In: *Geophy. J. Int.* 18.5 (1969), pp. 441–460. DOI: 10.1111/j.1365-246X.1969.tb03370.x

1970 •Hans Ramberg. "Folding of laterally compressed multilayers in the field of gravity, I". in: *Phys. Earth. Planet. Inter.* 2.4 (1970), pp. 203–232. DOI: 10.1016/0031-9201(70)90010-5

1971 •KE Torrance and DL Turcotte. "Thermal convection with large viscosity variations". In: *Journal of Fluid Mechanics* 47.1 (1971), pp. 113–125. DOI: 10.1017/S002211207100096X
•KE Torrance and DL Turcotte. "Structure of convection cells in the mantle". In: *J. Geophys. Res.: Solid Earth* 76.5 (1971), pp. 1154–1161. DOI: 10.1029/JB076i005p01154

1973 •Frank M Richter. "Convection and the large-scale circulation of the mantle". In: *J. Geophys. Res.: Solid Earth* 78.35 (1973), pp. 8735–8745. DOI: 10.1029/JB078i035p08735
•Frank M Richter. "Dynamical models for sea floor spreading". In: *Reviews of Geophysics* 11.2 (1973), pp. 223–287. DOI: 10.1029/RG011i002p00223
•Dan McKenzie, Jean Roberts, and Nigel Weiss. "Numerical models of convection in the earth's mantle". In: *Tectonophysics* 19.2 (1973), pp. 89–103. DOI: 10.1016/0040-1951(73)90034-6

1974 •Patrick Cassen and Ray T Reynolds. "Convection in the Moon: Effect of variable viscosity". In: *J. Geophys. Res.: Solid Earth* 79.20 (1974), pp. 2937–2944. DOI: 10.1029/JB079i020p02937
•Dan P McKenzie, Jean M Roberts, and Nigel O Weiss. "Convection in the Earth's mantle: towards a numerical simulation". In: *Journal of Fluid Mechanics* 62.3 (1974), pp. 465–538. DOI: 10.1017/S0022112074000784

1975 •J.M. Hewitt, D.P. McKenzie, and N.O. Weiss. "Dissipative heating in convective flows". In: *J. Fluid Mech.* 68.4 (1975), pp. 721–738. DOI: 10.1017/S002211207500119X
•EM Parmentier, DL Turcotte, and KE Torrance. "Numerical experiments on the structure of mantle plumes". In: *J. Geophys. Res.: Solid Earth* 80.32 (1975), pp. 4417–4424. DOI: 10.1029/JB080i032p04417

1976 •EM Parmentier, DL Turcotte, and KE Torrance. "Studies of finite amplitude non-Newtonian thermal convection with application to convection in the Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 81.11 (1976), pp. 1839–1846. DOI: 10.1029/JB081i011p01839

1977 •Geoffrey F Davies. "Whole-mantle convection and plate tectonics". In: *Geophy. J. Int.* 49.2 (1977), pp. 459–486. DOI: 10.1111/j.1365-246X.1977.tb03717.x
•Geoffrey F Davies. "Viscous mantle flow under moving lithospheric plates and under subduction zones". In: *Geophy. J. Int.* 49.3 (1977), pp. 557–563. DOI: 10.1111/j.1365-246X.1977.tb01303.x

1978 •Peter Bird. "Finite element modeling of lithosphere deformation: the Zagros collision orogeny". In: *Tectonophysics* 50.2-3 (1978), pp. 307–336
•Albert T Hsui. "Numerical simulation of finite-amplitude thermal convection with large viscosity variation in axisymmetric spherical geometry: effect of mechanical boundary conditions". In: *Tectonophysics* 50.2-3 (1978), pp. 147–162. DOI: 10.1016/0040-1951(78)90132-4
•D.K. Gartling. *Nachos - A finite element computer program for incompressible flow problems*. Tech. rep. Sand77-1333. Sandia Laboratories, 1978
•Aaron Tovish, Gerald Schubert, and Bruce P Luyendyk. "Mantle flow pressure and the angle of subduction: Non-Newtonian corner flows". In: *J. Geophys. Res.: Solid Earth* 83.B12 (1978), pp. 5892–5898
•W.-D. Woidt. "Finite element calculations applied to salt dome analysis". In: *Tectonophysics* 50 (1978), pp. 369–386. DOI: 10.1016/0040-1951(78)90143-9
•F. Richter and D. McKenzie. "Simple plate models of mantle convection". In: *J. Geophys.* 44 (1978), pp. 441–471. DOI: xxxx

1979 •D.F. Griffiths. "Finite Elements for Incompressible Flow". In: *Math. Meth. in the Appl. Sci.* 1 (1979), pp. 16–31
•Dan McKenzie. "Finite deformation during fluid flow". In: *Geophy. J. Int.* 58.3 (1979), pp. 689–715. DOI: 10.1111/j.1365-246X.1979.tb04803.x
•L. Bodri and B. Bodri. "Flow, stress and temperature in island arc areas". In: *Geophysical & Astrophysical Fluid Dynamics* 13.1 (1979), pp. 95–105. DOI: 10.1080/03091927908243763
•U Kopitzke et al. "Finite element convection models: comparison of shallow and deep mantle convection, and temperatures in the mantle". In: *Journal of Geophysics* 46.1 (1979), pp. 97–121. DOI: xxx
•Richard A Lux, Geoffrey F Davies, and John H Thomas. "Moving lithospheric plates and mantle convection". In: *Geophy. J. Int.* 58.1 (1979), pp. 209–228

1980 •Gary T. Jarvis and Dan P. McKenzie. "Convection in a compressible fluid with infinite Prandtl number". In: *Journal of Fluid Mechanics* 96.3 (1980), pp. 515–583. DOI: 10.1017/S002211208000225X
•Abdelfattah Zebib, Gerald Schubert, and Joe M Straus. "Infinite Prandtl number thermal convection in a spherical shell". In: *Journal of Fluid Mechanics* 97.2 (1980), pp. 257–277. DOI: 10.1017/S0022112080002558
•D.A. Yuen and W.R. Peltier. "Mantle plumes and the thermal stability of the D" layer". In: *Geophys. Res. Lett.* 7.9 (1980), pp. 625–628. DOI: 10.1029/GL007i009p00625

1981 •Wolfgang R Jacoby and Harro Schmeling. "Convection experiments and the driving mechanism". In: *Geologische Rundschau* 70.1 (1981), pp. 207–230. DOI: 10.1007/BF01764323

•Gary T Jarvis and WR Peltier. "Effects of lithospheric rigidity on ocean floor bathymetry and heat flow". In: *Geophys. Res. Lett.* 8.8 (1981), pp. 857–860. DOI: 10.1029/GL008i008p00857

•H. Schmeling and W.R. Jacoby. "On modelling the lithosphere in mantle convection with non-linear rheology". In: *Journal of Geophysics* 50 (1981), pp. 89–100

•D.A. Yuen, W.R. Peltier, and G. Schubert. "On the existence of a second scale of convection in the upper mantle". In: *Geophysical Journal of the Royal Astronomical Society* 65.1 (1981), pp. 171–190. DOI: 10.1111/j.1365-246X.1981.tb02707.x

•Frank M Richter and Dan P McKenzie. "Parameterizations for the horizontally averaged temperature of infinite Prandtl number convection". In: *J. Geophys. Res.: Solid Earth* 86.B3 (1981), pp. 1738–1744. DOI: 10.1029/JB086iB03p01738

•Greg A Houseman, D Po McKenzie, and Peter Molnar. "Convective instability of a thickened boundary layer and its relevance for the thermal evolution of continental convergent belts". In: *J. Geophys. Res.: Solid Earth* 86.B7 (1981), pp. 6115–6132. DOI: 10.1029/JB086iB07p06115

1982 •P. England. "Some numerical investigations of large scale continental deformation". In: *Mountain Building Processes*. Academic Press, 1982, pp. 129–189. DOI: xxxx

•Gary T Jarvis and WR Peltier. "Mantle convection as a boundary layer phenomenon". In: *Geophy. J. Int.* 68.2 (1982), pp. 389–427. DOI: 10.1111/j.1365-246X.1982.tb04907.x

•Wolfgang R Jacoby and Harro Schmeling. "On the effects of the lithosphere on mantle convection and evolution". In: *Phys. Earth. Planet. Inter.* 29.3-4 (1982), pp. 305–319. DOI: 10.1016/0031-9201(82)90019-X

•Frank J Spera, David A Yuen, and Stephen J Kirschvink. "Thermal boundary layer convection in silicic magma chambers: Effects of temperature-dependent rheology and implications for thermogravitational chemical fractionation". In: *J. Geophys. Res.: Solid Earth* 87.B10 (1982), pp. 8755–8767. DOI: 10.1029/JB087iB10p08755

1983 •Takeshi Matsumoto and Yoshibumi Tomoda. "Numerical simulation of the initiation of subduction at the fracture zone". In: *Journal of Physics of the Earth* 31.3 (1983), pp. 183–194. DOI: 10.4294/jpe1952.31.183

•U Christensen. "Convection in a variable-viscosity fluid: Newtonian versus power-law rheology". In: *Earth Planet. Sci. Lett.* 64.1 (1983), pp. 153–162. DOI: 10.1016/0012-821X(83)90060-2

•Ulrich Christensen. "A numerical model of coupled subcontinental and oceanic convection". In: *Tectonophysics* 95.1-2 (1983), pp. 1–23. DOI: 10.1016/0040-1951(83)90256-1

1984 •U.R. Christensen and D.A. Yuen. "The interaction of a subducting lithospheric slab with a chemical or phase boundary". In: *J. Geophys. Res.: Solid Earth* 89(B6) (1984), pp. 4389–4402. DOI: 10.1029/JB089iB06p04389

•U Christensen. "Convection with pressure-and temperature-dependent non-Newtonian rheology". In: *Geophy. J. Int.* 77.2 (1984), pp. 343–384. DOI: 10.1111/j.1365-246X.1984.tb01939.x

•GT Jarvis. "Time-dependent convection in the Earth's mantle". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 305–327. DOI: 10.1016/0031-9201(84)90054-2

•U Hansen and A Ebel. "Experiments with a numerical model related to mantle convection: boundary layer behaviour of small-and large scale flows". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 374–390. DOI: 10.1016/0031-9201(84)90058-X

1985 •NRA Hoffman and DP McKenzie. "The destruction of geochemical heterogeneities by differential fluid motions during mantle convection". In: *Geophy. J. Int.* 82.2 (1985), pp. 163–206. DOI: 10.1111/j.1365-246X.1985.tb05134.x

•Ulrich R Christensen and David A Yuen. "Layered convection induced by phase transitions". In: *J. Geophys. Res.: Solid Earth* 90.B12 (1985), pp. 10291–10300. DOI: 10.1029/JB090iB12p10291

•JX Mitrovica and GT Jarvis. "Surface deflections due to transient subduction in a convecting mantle". In: *Tectonophysics* 120.3-4 (1985), pp. 211–237. DOI: 10.1016/0040-1951(85)90052-6

•Satoru Honda. "Thermal structure beneath Tohoku, northeast Japan". In: *Tectonophysics* 112.1-4 (1985), pp. 69–102. DOI: 10.1016/0040-1951(85)90173-8

1986 •Michael Gurnis and Geoffrey F Davies. "Mixing in numerical models of mantle convection incorporating plate kinematics". In: *J. Geophys. Res.: Solid Earth* 91.B6 (1986), pp. 6375–6395. DOI: 10.1029/JB091iB06p06375

•Claire Harvey Craig and Dan McKenzie. "The existence of a thin low-viscosity layer beneath the lithosphere". In: *Earth Planet. Sci. Lett.* 78.4 (1986), pp. 420–426

•Geoffrey F Davies and Michael Gurnis. "Interaction of mantle dregs with convection: Lateral heterogeneity at the core-mantle boundary". In: *Geophys. Res. Lett.* 13.13 (1986), pp. 1517–1520

•Geoffrey F Davies. "Mantle convection under simulated plates: effects of heating modes and ridge and trench migration, and implications for the core-mantle boundary, bathymetry, the geoid and Benioff zones". In: *Geophys. J. R. astr. Soc* 84.1 (1986), pp. 153–183

•GT Jarvis and WR Peltier. "Lateral heterogeneity in the convecting mantle". In: *J. Geophys. Res.: Solid Earth* 91.B1 (1986), pp. 435–451. DOI: 10.1029/JB091iB01p00435

•W Roger Buck. "Small-scale convection induced by passive rifting: the cause for uplift of rift shoulders". In: *Earth Planet. Sci. Lett.* 77.3-4 (1986), pp. 362–372. DOI: 10.1016/0012-821X(86)90146-9

1987 •Ulrich R Christensen. "Time-dependent convection in elongated Rayleigh-Benard cells". In: *Geophys. Res. Lett.* 14.3 (1987), pp. 220–223. DOI: 10.1029/GL014i003p00220

•Harro Schmeling. "On the relation between initial conditions and late stages of Rayleigh-Taylor instabilities". In: *Tectonophysics* 133.1-2 (1987), pp. 65–80. DOI: 10.1016/0040-1951(87)90281-2

•J. Revenaugh and B. Parsons. "Dynamic topography and gravity anomalies for fluid layers whose viscosity varies exponentially with depth". In: *Geophysical Journal of the Royal Astronomical Society* 90.2 (1987), pp. 349–368. DOI: 10.1111/j.1365-246X.1987.tb00731.x

•Charles G Speziale. "On the advantages of the vorticity-velocity formulation of the equations of fluid dynamics". In: *J. Comp. Phys.* 73 (1987), pp. 476–480

•Peter Olson. "A comparison of heat transfer laws for mantle convection at very high Rayleigh numbers". In: *Phys. Earth. Planet. Inter.* 48.1-2 (1987), pp. 153–160. DOI: 10.1016/0031-9201(87)90118-X

•Marc Spiegelman and Dan McKenzie. "Simple 2-D models for melt extraction at mid-ocean ridges and island arcs". In: *Earth Planet. Sci. Lett.* 83.1-4 (1987), pp. 137–152

• JX Mitrovica and GT Jarvis. "A numerical study of thermal convection between rigid horizontal boundaries". In: *Geophysical & Astrophysical Fluid Dynamics* 38.3 (1987), pp. 193–223. DOI: 10.1080/03091928708219204

• U. Hansen and D.A. Yuen. "Evolutionary structures in double-diffusive convection in magma chambers". In: *Geophys. Res. Lett.* 14.11 (1987), pp. 1099–1102. DOI: 10.1029/GL014i011p01099

1988 • MGG Foreman and AF Bennett. "On no-slip boundary conditions for the incompressible Navier-Stokes equations". In: *Dynamics of atmospheres and oceans* 12.1 (1988), pp. 47–70. DOI: 10.1016/0377-0265(88)90014-0

• A.P. Vincent and D.A. Yuen. "Thermal attractor in chaotic convection with high-Prandtl-number fluids". In: *Physical Review A* 38.1 (1988), pp. 328–334. DOI: 10.1103/PhysRevA.38.328

• Max D Gunzburger and Janet S Peterson. "On finite element approximations of the streamfunction-vorticity and velocity-vorticity equations". In: *Int. J. Num. Meth. Fluids* 8.10 (1988), pp. 1229–1240. DOI: 10.1002/fld.1650081010

• G.F. Davies. "Role of the lithosphere in mantle convection". In: *J. Geophys. Res.: Solid Earth* 93.B9 (1988), pp. 10451–10466. DOI: 10.1029/JB093iB09p10451

• Ulrich Hansen and Adolf Ebel. "Time-dependent thermal convection-a possible explanation for a multiscale flow in the Earth's mantle". In: *Geophy. J. Int.* 94.2 (1988), pp. 181–191. DOI: 10.1111/j.1365-246X.1988.tb05895.x

1989 • U.R. Christensen and D.A. Yuen. "Time-dependent convection with non-Newtonian viscosity". In: *J. Geophys. Res.: Solid Earth* 94.B1 (1989), pp. 814–820. DOI: 10.1029/JB094iB01p00814

• G.F. Davies. "Mantle convection model with a dynamic plate: topography, heat flow and gravity anomalies". In: *Geophy. J. Int.* 98.3 (1989), pp. 461–464. DOI: 10.1111/j.1365-246X.1989.tb02283.x

• S.A. Weinstein, P.L. Olson, and D.A. Yuen. "Time-dependent large aspect-ratio thermal convection in the earth's mantle". In: *Geophysical & Astrophysical Fluid Dynamics* 47.1-4 (1989), pp. 157–197. DOI: 10.1080/03091928908221820

• Ulrich Hansen and David A Yuen. "Dynamical influences from thermal-chemical instabilities at the core-mantle boundary". In: *Geophys. Res. Lett.* 16.7 (1989), pp. 629–632. DOI: 10.1029/GL016i007p00629

• Harro Schmeling. "Compressible convection with constant and variable viscosity: The effect on slab formation, geoid, and topography". In: *J. Geophys. Res.: Solid Earth* 94.B9 (1989), pp. 12463–12481

• Volker Steinbach, Ulrich Hansen, and Adolf Ebel. "Compressible convection in the earth's mantle: a comparison of different approaches". In: *Geophys. Res. Lett.* 16.7 (1989), pp. 633–636. DOI: 10.1029/GL016i007p00633

• P. Machetel and D.A. Yuen. "Penetrative convective flows induced by internal heating and mantle compressibility". In: *J. Geophys. Res.: Solid Earth* 94.B8 (1989), pp. 10, 609–10, 626. DOI: 10.1029/JB094iB08p10609

1990 • GA Houseman. "The thermal structure of mantle plumes: axisymmetric or triple-junction?" In: *Geophy. J. Int.* 102.1 (1990), pp. 15–24. DOI: 10.1111/j.1365-246X.1990.tb00527.x

• GA Houseman. "Boundary conditions and efficient solution algorithms for the potential function formulation of the 3-D viscous flow equations". In: *Geophy. J. Int.* 100.1 (1990), pp. 33–38. DOI: 10.1111/j.1365-246X.1990.tb04565.x

• LP Solheim and WR Peltier. "Heat transfer and the onset of chaos in a spherical, axisymmetric, anelastic model of whole mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 53.4 (1990), pp. 205–255

• B.J. Travis et al. "A benchmark comparison of numerical methods for infinite Prandtl number thermal convection in two-dimensional Cartesian geometry". In: *Geophysical & Astrophysical Fluid Dynamics* 55.3-4 (1990), pp. 137–160

• Bryan Travis, Peter Olson, and Gerald Schubert. "The transition from two-dimensional to three-dimensional planforms in infinite-Prandtl-number thermal convection". In: *Journal of Fluid Mechanics* 216 (1990), pp. 71–91. DOI: 10.1017/S0022112090000349

• GT Jarvis and WR Peltier. "Low-wavenumber signatures of time-dependent mantle convection". In: *Phys. Earth. Planet. Inter.* 59.3 (1990), pp. 182–194. DOI: 10.1016/0031-9201(90)90224-L

• U. Hansen and D.A. Yuen. "Nonlinear physics of double-diffusive convection in geological systems". In: *Earth Science Reviews* 29.1-4 (1990), pp. 385–399. DOI: 10.1016/0012-8252(90)90050-6

1991 • William SD Wilcock and JA Whitehead. "The Rayleigh-Taylor instability of an embedded layer of low-viscosity fluid". In: *J. Geophys. Res.: Solid Earth* 96.B7 (1991), pp. 12193–12200. DOI: 10.1029/91JB00339

• Philip M Gresho. "Some current CFD issues relevant to the incompressible Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 87.2-3 (1991), pp. 201–252. DOI: 10.1016/0045-7825(91)90006-R

• A.V. Malevsky and D.A. Yuen. "Characteristics-based methods applied to infinite Prandtl number thermal convection in the hard turbulent regime". In: *Physics of Fluids A* 3.9 (1991), pp. 2105–2115. DOI: 10.1063/1.857893

• Ulrich Hansen, David A Yuen, and Sherri E Kroening. "Effects of depth-dependent thermal expansivity on mantle circulations and lateral thermal anomalies". In: *Geophys. Res. Lett.* 18.7 (1991), pp. 1261–1264. DOI: 10.1029/91GL01288

• Louise H Kellogg and Cheryl A Stewart. "Mixing by chaotic convection in an infinite Prandtl number fluid and implications for mantle convection". In: *Physics of Fluids A: Fluid Dynamics* 3.5 (1991), pp. 1374–1378. DOI: 10.1063/1.858067

• GT Jarvis. "Two-dimensional numerical models of mantle convection". In: *Advances in geophysics*. Vol. 33. 1991, pp. 1–80. DOI: 10.1016/S0065-2687(08)60440-9

1992 • R.R. Christensen. "An Eulerian technique for thermomechanical modeling of lithospheric extension". In: *J. Geophys. Res.: Solid Earth* 97.B2 (1992), pp. 2015–2036. DOI: 10.1029/91JB02642

• U. Hansen, D.A. Yuen, and S.E. Kroening. "Mass and Heat Transport in Strongly Time-Dependent Thermal Convection at Infinite Prandtl Number". In: *Geophysical & Astrophysical Fluid Dynamics* 63.1-4 (1992), pp. 67–89. DOI: 10.1080/03091929208228278

• U. Hansen, D.A. Yuen, and A.V. Malevsky. "Comparison of steady-state and strongly chaotic thermal convection at high Rayleigh number". In: *Physical Review A* 46.8 (1992), pp. 4742–4754. DOI: 10.1103/PhysRevA.46.4742

• Andrei V Malevsky, David A Yuen, and LM Weyer. "Viscosity and thermal fields associated with strongly chaotic non-Newtonian thermal convection". In: *Geophys. Res. Lett.* 19.2 (1992), pp. 127–130

• A. Poliakov and Y. Podlachikov. "Diapirism and topography". In: *Geophy. J. Int.* 109 (1992), pp. 553–564. DOI: 10.1111/j.1365-246X.1992.tb00117.x

• P. van Keken, D.A. Yuen, and A. van den Berg. "Pulsating diapiric flows: Consequences of vertical variations in mantle creep laws". In: *Earth Planet. Sci. Lett.* 112 (1992), pp. 179–194. DOI: 10.1016/0012-821X(92)90015-N

• R.F. Weinberg and H. Schmeling. "Polydiapirs: multiwavelength gravity structures". In: *Journal of Structural Geology* 14.4 (1992), pp. 425–436. DOI: 10.1016/0191-8141(92)90103-4

• A.V. Malevsky and D.A. Yuen. "Strongly chaotic non-newtonian mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 65.1-4 (1992), pp. 149–171. DOI: 10.1080/03091929208225244

1149

1993 •Ulrich Hansen, David A Yuen, SE Kroening, and TB Larsen. "Dynamical consequences of depth-dependent thermal expansivity and viscosity on mantle circulations and thermal structure". In: *Phys. Earth. Planet. Inter.* 77.3-4 (1993), pp. 205–223. DOI: `10.1016/0031-9201(93)90099-U`

•U. Hansen and D.A. Yuen. "High Rayleigh number regime of temperature-dependent viscosity convection and the Earth's early thermal history". In: *Geophys. Res. Lett.* 20.20 (1993), pp. 2191–2194. DOI: `10.1029/93GL02416`

•A.V. Malevsky and D.A. Yuen. "Plume structures in the hard-turbulent regime of three-dimensional infinite Prandtl number convection". In: *Geophys. Res. Lett.* 20.5 (1993), pp. 383–386. DOI: `10.1029/93GL00293`

•Arie P van den Berg, Peter E van Keken, and David A Yuen. "The effects of a composite non-Newtonian and Newtonian rheology on mantle convection". In: *Geophy. J. Int.* 115.1 (1993), pp. 62–78. DOI: `10.1111/j.1365-246X.1993.tb05588.x`

•David Bercovici. "A simple model of plate generation from mantle flow". In: *Geophy. J. Int.* 114.3 (1993), pp. 635–650. DOI: `10.1111/j.1365-246X.1993.tb06993.x`

•G Guj and F Stella. "A vorticity-velocity method for the numerical solution of 3D incompressible flows". In: *J. Comp. Phys.* 106.2 (1993), pp. 286–298. DOI: `10.1016/S0021-9991(83)71108-3`

•G.T. Jarvis. "Effects of curvature on two-dimensional models of mantle convection: cylindrical polar coordinates". In: *J. Geophys. Res.: Solid Earth* 98.B3 (1993), pp. 4477–4485. DOI: `10.1029/92JB02117`

•Julian P Lowman and Gary T Jarvis. "Mantle convection flow reversals due to continental collisions". In: *Geophys. Res. Lett.* 20.19 (1993), pp. 2087–2090. DOI: `10.1029/93GL02047`

•Volker Steinbach, David A Yuen, and Wuling Zhao. "Instabilities from phase transitions and the timescales of mantle thermal evolution". In: *Geophys. Res. Lett.* 20.12 (1993), pp. 1119–1122. DOI: `10.1029/93GL01243`

•P.E. van Keken, C.J. Spiers, A.P. van den Berg, and E.J. Muyzert. "The effective viscosity of rocksalt: implementation of steady-state creep laws in numerical models of salt diapirism". In: *Tectonophysics* 225 (1993), pp. 457–476. DOI: `10.1016/0040-1951(93)90310-G`

•P.E. van Keken, D.A. Yuen, and A.P. van den Berg. "The effects of shallow rheological boundaries in the upper mantle on inducing shorter time scales of diapiric flows". In: *Geophys. Res. Lett.* 20.18 (1993), pp. 1927–1930. DOI: `10.1029/93GL01768`

•Yoshitsugu Furukawa. "Depth of the decoupling plate interface and thermal structure under arcs". In: *J. Geophys. Res.: Solid Earth* 98.B11 (1993), pp. 20005–20013

1994 •F.H. Busse et al. "3D convection at infinite Prandtl number in Cartesian geometry - a benchmark comparison". In: *Geophysical & Astrophysical Fluid Dynamics* 75.1 (1994), pp. 39–59. DOI: `10.1080/03091929408203646`

•Larry P Solheim and WR Peltier. "Avalanche effects in phase transition modulated thermal convection: A model of Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 99.B4 (1994), pp. 6997–7018. DOI: `10.1029/93JB02168`

•Tomoeki Nakakuki, Hiroki Sato, and Hiromi Fujimoto. "Interaction of the upwelling plume with the phase and chemical boundary at the 670 km discontinuity: Effects of temperature-dependent viscosity". In: *Earth Planet. Sci. Lett.* 121.3–4 (1994), pp. 369–384. DOI: `10.1016/0012-821X(94)90078-7`

•B.A. Buffet, C.W. Gable, and R.J. R.J. O'Connell. "Linear stability of a layered fluid with mobile surface plates". In: *J. Geophys. Res.: Solid Earth* 99(B10) (1994), pp. 19, 885–19, 900

•J. Schmalzl and U. Hansen. "Mixing the Earth's mantle by thermal convection: A scale dependent phenomenon". In: *Geophys. Res. Lett.* 21.11 (1994), pp. 987–990. DOI: `10.1029/94GL00049`

•N.J. Vlaar, P.E. van Keken, and A.P. van den Berg. "Cooling of the Earth in the Archean: Consequences of pressure-release melting in a hotter mantle". In: *Earth Planet. Sci. Lett.* 121 (1994), pp. 1–18. DOI: `10.1016/0012-821X(94)90028-0`

•Bryan Travis and Peter Olson. "Convection with internal heat sources and thermal turbulence in the Earth's mantle". In: *Geophy. J. Int.* 118.1 (1994), pp. 1–19. DOI: `10.1111/j.1365-246X.1994.tb04671.x`

•David A Yuen, DM Reuteler, S Balachandar, V Steinbach, AV Malevsky, and JJ Smedsmo. "Various influences on three-dimensional mantle convection with phase transitions". In: *Phys. Earth. Planet. Inter.* 86.1-3 (1994), pp. 185–203. DOI: `10.1016/0031-9201(94)05068-6`

•S. Zhong and M. Gurnis. "Role of plates and temperature-dependent viscosity in phase change dynamics". In: *J. Geophys. Res.: Solid Earth* 99.B8 (1994), p. 15903. DOI: `10.1029/94JB00545`

•U. Hansen and D.A. Yuen. "Effects of depth-dependent thermal expansivity on the interaction of thermal-chemical plumes with a compositional boundary". In: *Phys. Earth. Planet. Inter.* 86.1-3 (1994), pp. 205–221. DOI: `10.1016/0031-9201(94)05069-4`

1995 •J. Schmalzl, G.A. Houseman, and U. Hansen. "Mixing properties of three-dimensional (3-D) stationary convection". In: *Physics of Fluids* 7.5 (1995), pp. 1027–1033. DOI: `10.1063/1.868614`

•D. Bittner and H. Schmeling. "Numerical modelling of melting processes and induced diapirism in the lower crust". In: *Geophy. J. Int.* 123 (1995), pp. 59–70. DOI: `10.1111/j.1365-246X.1995.tb06661.x`

•NM Ribe, UR Christensen, and J Theissing. "The dynamics of plume-ridge interaction, 1: Ridge-centered plumes". In: *Earth Planet. Sci. Lett.* 134.1-2 (1995), pp. 155–168

•Volker Steinbach and David A Yuen. "The non-adiabatic nature of mantle convection as revealed by passive tracers". In: *Earth Planet. Sci. Lett.* 136.3-4 (1995), pp. 241–250. DOI: `10.1016/0012-821X(95)00166-A`

•SCR Dennis and JD Hudson. "Methods of solution of the velocity-vorticity formulation of the Navier-Stokes equations". In: *J. Comp. Phys.* 122.2 (1995), pp. 300–306

•Julian P Lowman and Gary T Jarvis. "Mantle convection models of continental collision and breakup incorporating finite thickness plates". In: *Phys. Earth. Planet. Inter.* 88.1 (1995), pp. 53–68

•VS Solomatov. "Scaling of temperature-and stress-dependent viscosity convection". In: *Physics of Fluids* 7.2 (1995), pp. 266–274

•Tine B Larsen, David A Yuen, and Andrei V Malevsky. "Dynamical consequences on fast subducting slabs from a self-regulating mechanism due to viscous heating in variable viscosity convection". In: *Geophys. Res. Lett.* 22.10 (1995), pp. 1277–1280. DOI: `10.1029/95GL01112`

1996 •T.B. Larsen, D.A. Yuen, A.V. Malevsky, and J.L. Smedsmo. "Dynamics of strongly time-dependent convection with non-Newtonian temperature-dependent viscosity". In: *Phys. Earth. Planet. Inter.* 94.1-2 (1996), pp. 75–103. DOI: `10.1016/0031-9201(95)03082-4`

•Terence D Barr and Gregory A Houseman. "Deformation fields around a fault embedded in a non-linear ductile medium". In: *Geophy. J. Int.* 125.2 (1996), pp. 473–490. DOI: `10.1111/j.1365-246X.1996.tb00012.x`

•David Bercovici. "Plate generation in a simple model of lithosphere-mantle flow with dynamic self-lubrication". In: *Earth Planet. Sci. Lett.* 144.1-2 (1996), pp. 41–51

•D Breuer, H Zhou, David A Yuen, and T Spohn. "Phase transitions in the Martian mantle: Implications for the planet's volcanic history". In: *J. Geophys. Res.: Planets* 101.E3 (1996), pp. 7531–7542. DOI: `10.1029/96JE00117`

•Ulrich R Christensen. "The influence of trench migration on slab penetration into the lower mantle". In: *Earth Planet. Sci. Lett.* 140.1-4 (1996), pp. 27–39. DOI: 10.1016/0012-821X(96)00023-4

•A. Lenardic and W. M. Kaula. "Near-surface thermal/chemical boundary layer convection at infinite Prandtl number: two-dimensional numerical experiments". In: *Geophy. J. Int.* 126.3 (1996), pp. 689–711. DOI: 10.1111/j.1365-246X.1996.tb04698.x

•Julian P Lowman and Gary T Jarvis. "Continental collisions in wide aspect ratio and high Rayleigh number two-dimensional mantle convection models". In: *J. Geophys. Res.: Solid Earth* 101.B11 (1996), pp. 25485–25497. DOI: 10.1029/96JB02568

•J.X. Mitrovica, R.N. Pysklywec, C. Beaumont, and A. Rutty. "The Devonian to Permian sedimentation of the Russian platform: An example of subduction-controlled long-wavelength tilting of continents". In: *Journal of Geodynamics* 22.1-2 (1996), pp. 79–96. DOI: 10.1016/0264-3707(96)00008-7

•NM Ribe. "The dynamics of plume-ridge interaction - II. Off-ridge plumes". In: *J. Geophys. Res.: Solid Earth* 101.B7 (1996), pp. 16195–16204

•J. Schmalzl, G.A. Houseman, and U. Hansen. "Mixing in vigorous, time-dependent three-dimensional convection and application to Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 101.B10 (1996), pp. 21847–21858

•Arie P van den Berg and David A Yuen. "Is the lower-mantle rheology Newtonian today?" In: *Geophys. Res. Lett.* 23.16 (1996), pp. 2033–2036. DOI: 10.1029/96GL02065

•S. Zhong, M. Gurnis, and L. Moresi. "Free-surface formulation of mantle convection-I. Basic theory and application to plumes". In: *Geophy. J. Int.* 127.3 (1996), pp. 708–718. DOI: 10.1111/j.1365-246X.1996.tb04049.x

•Shijie Zhong. "Analytic solutions for Stokes' flow with lateral variations in viscosity". In: *Geophy. J. Int.* 124.1 (1996), pp. 18–28. DOI: 10.1111/j.1365-246X.1996.tb06349.x

1997 •Tine B Larsen, David A Yuen, Jiří Moser, and Bengt Fornberg. "A high-order finite-difference method applied to large Rayleigh number mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 84.1-2 (1997), pp. 53–83. DOI: 10.1080/03091929708208973

•D. Olbertz, M.J.R. Wortel, and U. Hansen. "Trench migration and subduction zone geometry". In: *Geophys. Res. Lett.* 24.3 (1997), pp. 221–224. DOI: 10.1029/96GL03971

•Uwe Walzer and Roland Hendel. "Tectonic episodicity and convective feedback mechanisms". In: *Phys. Earth. Planet. Inter.* 100.1-4 (1997), pp. 167–188. DOI: 10.1016/S0031-9201(96)03238-4

•Doris Breuer, Dave A Yuen, and Tilman Spohn. "Phase transitions in the Martian mantle: Implications for partially layered convection". In: *Earth Planet. Sci. Lett.* 148.3-4 (1997), pp. 457–469. DOI: 10.1016/S0012-821X(97)00049-6

•Wanda DeLandro-Clarke and Gary T Jarvis. "Numerical models of mantle convection with secular cooling". In: *Geophy. J. Int.* 129.1 (1997), pp. 183–193. DOI: 10.1111/j.1365-246X.1997.tb00948.x

•T.B. Larsen and D.A. Yuen. "Ultrafast upwelling bursting through the upper mantle". In: *Earth Planet. Sci. Lett.* 146.3-4 (1997), pp. 393–399. DOI: 10.1016/S0012-821X(96)00247-6

•T.B. Larsen and D.A. Yuen. "Fast plumeheads: Temperature-dependent versus non-Newtonian rheology". In: *Geophys. Res. Lett.* 24.16 (1997), pp. 1995–1998. DOI: 10.1029/97GL01886

•P.E. van Keken, S.D. King, H. Schmeling, U.R. Christensen, D. Neumeister, and M.-P. Doin. "A comparison of methods for the modeling of thermochemical convection". In: *J. Geophys. Res.: Solid Earth* 102.B10 (1997), pp. 22, 477–22, 495

•A.P. Van Den Berg and D.A. Yuen. "The role of shear heating in lubricating mantle flow". In: *Earth Planet. Sci. Lett.* 151.1-2 (1997), pp. 33–42. DOI: 10.1016/S0012-821X(97)00110-6

•Uwe Walzer and Roland Hendel. "Tectonic episodicity and convective feedback mechanisms". In: *Phys. Earth. Planet. Inter.* 100.1-4 (1997), pp. 167–188. DOI: 10.1016/S0031-9201(96)03238-4

1998 •Helmut Harder. "Phase transitions and the three-dimensional planform of thermal convection in the Martian mantle". In: *J. Geophys. Res.: Planets* 103.E7 (1998), pp. 16775–16797. DOI: 10.1029/98JE01543

•B Schott and H Schmeling. "Delamination and detachment of a lithospheric root". In: *Tectonophysics* 296.3-4 (1998), pp. 225–247. DOI: 10.1016/S0040-1951(98)00154-1

•JA Gil and Maria José Jurado. "Geological interpretation and numerical modelling of salt movement in the Barbastro–Balaguer anticline, southern Pyrenees". In: *Tectonophysics* 293.3-4 (1998), pp. 141–155

•Boris M Naimark, Ali T Ismail-Zadeh, and Wolfgang R Jacoby. "Numerical approach to problems of gravitational instability of geostructures with advected material boundaries". In: *Geophy. J. Int.* 134.2 (1998), pp. 473–483. DOI: 10.1111/j.1365-246X.1998.tb07140.x

1999 •Uwe Walzer and Roland Hendel. "A new convection–fractionation model for the evolution of the principal geochemical reservoirs of the Earth's mantle". In: *Phys. Earth. Planet. Inter.* 112.3-4 (1999), pp. 211–256. DOI: 10.1016/S0031-9201(99)00035-7

•C Sotin and S Labrosse. "Three-dimensional thermal convection in an iso-viscous, infinite Prandtl number fluid heated from within and from below: applications to the transfer of heat through planetary mantles". In: *Phys. Earth. Planet. Inter.* 112.3-4 (1999), pp. 171–190. DOI: 10.1016/S0031-9201(99)00004-7

2008 •Juan-Luis Valera, Ana-María Negredo, and Antonio Villaseñor. "Asymmetric delamination and convective removal numerical modeling: comparison with evolutionary models for the Alboran Sea region". In: *Pure Appl. Geophys.* 165 (2008), pp. 1683–1706. DOI: 10.1007/s00024-008-0395-8

2011 •JL Valera, Ana M Negredo, and Ivone Jiménez-Munt. "Deep and near-surface consequences of root removal by asymmetric continental delamination". In: *Tectonophysics* 502.1-2 (2011), pp. 257–265. DOI: 10.1016/j.tecto.2010.04.002

Note that the StreamV code by Samuel relies on the Stream Function approach.

Chapter 22

Elasticity: physics, formulations and FEM

Let us start by clarifying notations:

variable name	symbol	unit
full stress tensor	$\boldsymbol{\sigma}$	Pa
deviatoric stress tensor	$\boldsymbol{\tau}$	Pa
strain tensor	$\boldsymbol{\varepsilon}$	-
elastic strain tensor	$\boldsymbol{\varepsilon}_{\mathrm{e}}$	-
visco-plastic strain tensor	$\boldsymbol{\varepsilon}_{\mathrm{vp}}$	-
total strain tensor	$\boldsymbol{\varepsilon}_{\mathrm{T}}$	-
strain rate tensor	$\dot{\boldsymbol{\varepsilon}}$	s^{-1}
Lamé parameter	λ	Pa
Shear modulus	μ	Pa
Bulk modulus	K	Pa
Poisson ratio	ν	-
Young's modulus	E	Pa
viscosity	η	Pa s
displacement	$\vec{\upsilon}$	m
velocity	$\vec{\nu}$	$\mathrm{m\,s}^{-1}$

What follows is a compilation of various sources, such as the Becker & Kaus lecture notes [71], the excellent paper by Beuchert and Podladchikov [92] (2010), the syllabus of R. Hassani [**XX**] and various books such as Sadd [1168].

One will find in the literature either 'elasto-viscosity' or 'visco-elasticity'. In what follows I have adopted the former notation with the acronym EV.

Once the equations have been laid out, one must then make a fundamental choice with regards to the type of code/calculations in the case of elasto-viscous rheologies: will the primary variable be displacement $\vec{\upsilon}$ or velocity $\vec{\nu}$? The latter is the common approach in the geodynamics community. The vast majority of codes are fluid flow solvers, formulated in velocity and pressure. Elasticity is usually added to such codes way after they were first used/written (eg ELEFANT, ASPECT, ...). For a purely elastic code displacement is the meaningful primary variable since the stress is formulated as a function of strain (not strain rate).

22.1 Basic equations

The strong form of the PDE that governs force balance in a medium is given by

$$\vec{\nabla} \cdot \boldsymbol{\sigma} + \vec{f} = \vec{0}$$

where $\boldsymbol{\sigma}$ is the full stress tensor and $\vec{f}$ is a body force (typically $\rho\vec{g}$).

The stress tensor is related to the strain tensor through the generalised Hooke's law[1]:

$$\sigma_{ij} = \sum_{kl} C_{ijkl}\varepsilon_{kl} \qquad \text{or} \qquad \boldsymbol{\sigma} = \boldsymbol{C} : \boldsymbol{\varepsilon} \tag{22.1}$$

where $\boldsymbol{C}$ is the fourth-order elastic tensor (which contains $3^4 = 81$ coefficients). The strain tensor is related to the displacement $\vec{\upsilon}$ as follows:

$$\boldsymbol{\varepsilon}(\vec{\upsilon}) = \frac{1}{2}(\vec{\nabla}\vec{\upsilon} + (\vec{\nabla}\vec{\upsilon})^T) \tag{22.2}$$

Due to the inherent symmetries of $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{C}$, only 21 elastic coefficients of the latter are independent. For isotropic linear media (which have the same physical properties in any direction), $\boldsymbol{C}$ can be reduced to only two independent numbers (for example the bulk modulus K and the shear modulus[2] μ that quantify the material's resistance to changes in volume and to shearing deformations, respectively). We find that

$$C_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$

so that Eq. (22.1) becomes:

$$\sigma_{ij} = \lambda\varepsilon_{kk}\delta_{ij} + 2\mu\varepsilon_{ij}$$

or

$$
\begin{aligned}
\boldsymbol{\sigma} &= \lambda\mathrm{Tr}[\boldsymbol{\varepsilon}(\vec{\upsilon})]\mathbf{1} + 2\mu\boldsymbol{\varepsilon}(\vec{\upsilon}) \\
&= \lambda(\vec{\nabla} \cdot \vec{\upsilon})\mathbf{1} + 2\mu\boldsymbol{\varepsilon}(\vec{\upsilon})
\end{aligned} \tag{22.3}
$$

where λ is the Lamé parameter and μ is the shear modulus. The term $\vec{\nabla}\cdot\vec{\upsilon} = \mathrm{Tr}[\boldsymbol{\varepsilon}(\vec{\upsilon})]$ is the isotropic dilation.

Very explicitly, and since the stress and strain tensors are symmetric, we have

$$
\begin{aligned}
\sigma_{xx} &= (\lambda + 2\mu)\varepsilon_{xx} + \lambda\varepsilon_{yy} + \lambda\varepsilon_{zz} \\
\sigma_{yy} &= \lambda\varepsilon_{xx} + (\lambda + 2\mu)\varepsilon_{yy} + \lambda\varepsilon_{zz} \\
\sigma_{zz} &= \lambda\varepsilon_{xx} + \lambda\varepsilon_{yy} + (\lambda + 2\mu)\varepsilon_{zz} \\
\sigma_{xy} &= 2\mu\varepsilon_{xy} \\
\sigma_{xz} &= 2\mu\varepsilon_{xz} \\
\sigma_{yz} &= 2\mu\varepsilon_{yz}
\end{aligned} \tag{22.4}
$$

This can be re-written in the 6-dimensional stress/strain space as

$$
\underbrace{\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix}}_{\vec{\sigma}} = \underbrace{\begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}}_{\boldsymbol{D}} \cdot \underbrace{\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ 2\varepsilon_{xy} \\ 2\varepsilon_{xz} \\ 2\varepsilon_{yz} \end{pmatrix}}_{\vec{\varepsilon}} \tag{22.5}
$$

---

[1] https://en.wikipedia.org/wiki/Hooke's_law
[2] It is also sometimes written G

Let us define the matrices

$$
\mathbf{\Lambda} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}
\qquad
\mathbf{\Xi} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

so that

$$
\boldsymbol{D} = \lambda\mathbf{\Xi} + \mu\mathbf{\Lambda}
$$

If we define the Young's modulus E and the Poisson's ratio as

$$
E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} \qquad \text{or} \qquad \nu = \frac{\lambda}{2(\lambda + \mu)} \tag{22.6}
$$

Then

$$
1 - \nu = 1 - \frac{\lambda}{2(\lambda + \mu)} = \frac{2(\lambda + \mu)}{2(\lambda + \mu)} - \frac{\lambda}{2(\lambda + \mu)} = \frac{\lambda + 2\mu}{2(\lambda + \mu)}
$$

$$
1 + \nu = 1 + \frac{\lambda}{2(\lambda + \mu)} = \frac{2(\lambda + \mu)}{2(\lambda + \mu)} + \frac{\lambda}{2(\lambda + \mu)} = \frac{3\lambda + 2\mu}{2(\lambda + \mu)} = \frac{E}{2\mu}
$$

$$
1 - 2\nu = 1 - 2\frac{\lambda}{2(\lambda + \mu)} = \frac{2(\lambda + \mu)}{2(\lambda + \mu)} - \frac{2\lambda}{2(\lambda + \mu)} = \frac{\mu}{\lambda + \mu}
$$

$$
\frac{E}{(1 + \nu)(1 - 2\nu)}(1 - \nu) = E\frac{\lambda + 2\mu}{2(\lambda + \mu)} \cdot \frac{2\mu}{E} \cdot \frac{\lambda + \mu}{\mu} = \lambda + 2\mu
$$

$$
\frac{E}{(1 + \nu)(1 - 2\nu)}\frac{1}{2}(1 - 2\nu) = E\frac{\mu}{\lambda + \mu} \cdot \frac{2\mu}{E} \cdot \frac{1}{2}\frac{\lambda + \mu}{\mu} = \mu
$$

$$
\frac{E}{(1 + \nu)(1 - 2\nu)}\nu = E\frac{\lambda}{2(\lambda + \mu)} \cdot \frac{2\mu}{E} \cdot \frac{\lambda + \mu}{\mu} = \lambda
$$

and in the end:

$$
\boldsymbol{D}_{\mathrm{3D}} = \begin{pmatrix} \lambda{+}2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda{+}2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda{+}2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} = \frac{E}{(1{+}\nu)(1{-}2\nu)} \begin{pmatrix} 1{-}\nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1{-}\nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1{-}\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1{-}2\nu}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1{-}2\nu}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1{-}2\nu}{2} \end{pmatrix} \tag{22.7}
$$

This matrix is the same as Eq. (3.6) on page 294 of Braess [136]. It is SPD for $0 \le \nu < 1/2$.

In terms of the compliance matrix $\boldsymbol{D}^{-1}$,

$$
\vec{\varepsilon} = \boldsymbol{D}^{-1} \cdot \vec{\sigma}
$$

with check these!

$$\boldsymbol{D}^{-1} = \frac{1}{\mu(3\lambda + 2\mu)} \begin{pmatrix} \lambda + \mu & -\lambda/2 & -\lambda/2 & 0 & 0 & 0 \\ -\lambda/2 & \lambda + \mu & -\lambda/2 & 0 & 0 & 0 \\ -\lambda/2 & -\lambda/2 & \lambda + \mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 3\lambda + 2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 3\lambda + 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 3\lambda + 2\mu \end{pmatrix}$$

then

$$\boldsymbol{D}^{-1} = \frac{1}{E} \begin{pmatrix} 1 & -\nu & -\nu & 0 & 0 & 0 \\ -\nu & 1 & -\nu & 0 & 0 & 0 \\ -\nu & -\nu & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2(1 + \nu) & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(1 + \nu) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(1 + \nu) \end{pmatrix}$$

Note that the determinant of $\boldsymbol{D}^{-1}$ is $8(1 + \nu)^5(1 - 2\nu)E^{-6}$, so that when $\nu \to 1/2$ (incompressible material), the compliance matrix is singular and the stress cannot be given as a function of strain [862].

The above equation also leads to:

$$
\begin{array}{rcl}
E\varepsilon_{xx} & = & \sigma_{xx} - \nu(\sigma_{yy} + \sigma_{zz}) \\
E\varepsilon_{yy} & = & \sigma_{yy} - \nu(\sigma_{xx} + \sigma_{zz}) \\
E\varepsilon_{zz} & = & \sigma_{zz} - \nu(\sigma_{xx} + \sigma_{yy}) \\
E\varepsilon_{xy} & = & (1 + \nu)\sigma_{xy} \\
E\varepsilon_{xz} & = & (1 + \nu)\sigma_{xz} \\
E\varepsilon_{yz} & = & (1 + \nu)\sigma_{yz}
\end{array}
\tag{22.8}
$$

The incompressibility (or bulk modulus) K is defined as $p = -K\vec{\nabla} \cdot \vec{\upsilon}$ where p is the pressure with

$$
\begin{array}{rcl}
p & = & -\dfrac{1}{3}\mathrm{tr}(\boldsymbol{\sigma}) \\[2mm]
& = & -\dfrac{1}{3}[\lambda(\vec{\nabla} \cdot \vec{\upsilon})\,\mathrm{tr}(\mathbf{1}) + 2\mu\,\mathrm{tr}[\boldsymbol{\varepsilon}(\vec{\upsilon})]] \\[2mm]
& = & -\dfrac{1}{3}[\lambda(\vec{\nabla} \cdot \vec{\upsilon})3 + 2\mu(\vec{\nabla} \cdot \vec{\upsilon})] \\[2mm]
& = & -\left(\lambda + \dfrac{2}{3}\mu\right)(\vec{\nabla} \cdot \vec{\upsilon})
\end{array}
\tag{22.9}
$$

so that

$$p = -K\vec{\nabla} \cdot \vec{\upsilon} \qquad \text{with} \qquad K = \lambda + \frac{2}{3}\mu \qquad \text{and} \qquad \boldsymbol{\sigma} = -p\mathbf{1} + 2\mu\boldsymbol{\varepsilon}^d(\vec{\upsilon})$$

Remark. *Eq. (22.1) and (22.3) are analogous to the ones that one has to solve in the context of viscous flow using the penalty method. In this case λ is the penalty coefficient, $\vec{\upsilon}$ is the velocity, and μ is the dynamic viscosity.*

Remark. *Note that sometimes authors define $p = -\lambda\vec{\nabla} \cdot \vec{\upsilon}$ instead so that then $\boldsymbol{\sigma} = -p\mathbf{1} + 2\mu\boldsymbol{\varepsilon}(\vec{\upsilon})$ (to be very clear, strain tensor is not deviatoric), see for instance Sanan, May, Bollhöfer, and Schenk [1182] (2020) or Hansbo, Larson, and Larson [563] (2001).*

The Lamé parameter λ and the shear modulus μ are also linked to the Poisson ratio ν, and E, Young's modulus:

$$\lambda = \mu \frac{2\nu}{1 - 2\nu} = \frac{\nu E}{(1 + \nu)(1 - 2\nu)} \qquad \text{with} \qquad E = 2\mu(1 + \nu)$$

The shear modulus, expressed often in GPa, describes the material's response to shear stress. The Poisson ratio describes the response in the direction orthogonal to uniaxial stress. The Young's modulus[3], expressed in GPa, describes the material's strain response to uniaxial stress in the direction of this stress.

In the future we will also need to express the deviatoric part of a tensor as a function of the tensor itself, all in vector format. Let us consider the stress tensor. Then we have $\boldsymbol{\tau} = \mathrm{dev}(\boldsymbol{\sigma}) = \boldsymbol{\sigma} - \frac{1}{3}\mathrm{tr}[\boldsymbol{\sigma}]\mathbf{1}$ which becomes

$$\vec{\tau} = \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} - \frac{1}{3}\begin{pmatrix} \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\ \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\ \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{3}\sigma_{xx} - \frac{1}{3}\sigma_{yy} - \frac{1}{3}\sigma_{zz} \\ -\frac{1}{3}\sigma_{xx} + \frac{2}{3}\sigma_{yy} - \frac{1}{3}\sigma_{zz} \\ -\frac{1}{3}\sigma_{xx} - \frac{1}{3}\sigma_{yy} + \frac{2}{3}\sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\tilde{\boldsymbol{\Lambda}}^d} \cdot \vec{\sigma}$$

or,

$$\vec{\tau} = \tilde{\boldsymbol{\Lambda}}^d \cdot \vec{\sigma} \tag{22.10}$$

where $\tilde{\boldsymbol{\Lambda}}^d$ is a deviatoric projection matrix.

Using the definition of K above, we have $\lambda + 2\mu = K + \frac{4}{3}\mu$ and $\lambda = K - \frac{2}{3}\mu$ so that the $\boldsymbol{D}$ matrix can also be written as a function of K, μ:

$$\begin{aligned}
\boldsymbol{D}_{3\mathrm{D}} &= \begin{pmatrix} K + \frac{4}{3}\mu & K - \frac{2}{3}\mu & K - \frac{2}{3}\mu & 0 & 0 & 0 \\ K - \frac{2}{3}\mu & K + \frac{4}{3}\mu & K - \frac{2}{3}\mu & 0 & 0 & 0 \\ K - \frac{2}{3}\mu & K - \frac{2}{3}\mu & K + \frac{4}{3}\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \\[2mm]
&= \begin{pmatrix} K & K & K & 0 & 0 & 0 \\ K & K & K & 0 & 0 & 0 \\ K & K & K & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{4}{3}\mu & -\frac{2}{3}\mu & -\frac{2}{3}\mu & 0 & 0 & 0 \\ -\frac{2}{3}\mu & \frac{4}{3}\mu & -\frac{2}{3}\mu & 0 & 0 & 0 \\ -\frac{2}{3}\mu & -\frac{2}{3}\mu & \frac{4}{3}\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \\[2mm]
&= K\boldsymbol{\Xi} + \mu\boldsymbol{\Lambda}^d \tag{22.11}
\end{aligned}$$

$$\boldsymbol{D}_{3\mathrm{D}} = K\boldsymbol{\Xi} + \mu\boldsymbol{\Lambda}^d$$

The expression above is to be found at `https://en.wikipedia.org/wiki/Linear_elasticity`

---

[3]`https://en.wikipedia.org/wiki/Young's_modulus`

22.2   Plane strain

Typically one of the spatial dimensions (e.g. z) is very large compared to the other two. As a consequence displacements υ_z and displacement derivatives ∂_z in the z-direction are assumed to be negligible, i.e. $\varepsilon_{zz} = \varepsilon_{xz} = \varepsilon_{yz} = 0$ and Eqs. (22.8) become:

$$
\begin{aligned}
E\varepsilon_{xx} &= \sigma_{xx} - \nu(\sigma_{yy} + \sigma_{zz}) \\
E\varepsilon_{yy} &= \sigma_{yy} - \nu(\sigma_{xx} + \sigma_{zz}) \\
E\varepsilon_{zz} &= \sigma_{zz} - \nu(\sigma_{xx} + \sigma_{yy}) \\
E\varepsilon_{xy} &= (1+\nu)\sigma_{xy} \\
E\varepsilon_{xz} &= (1+\nu)\sigma_{xz} \\
E\varepsilon_{yz} &= (1+\nu)\sigma_{yz}
\end{aligned}
$$

leading to $\sigma_{xz} = \sigma_{yz} = 0$, $\sigma_{zz} = \nu(\sigma_{xx} + \sigma_{yy})$ and

$$
\begin{aligned}
E\varepsilon_{xx} &= \sigma_{xx} - \nu(\sigma_{yy} + \sigma_{zz}) \\
&= \sigma_{xx} - \nu(\sigma_{yy} + \nu(\sigma_{xx} + \sigma_{yy})) \\
&= (1-\nu^2)\sigma_{xx} - \nu(1+\nu)\sigma_{yy} \\
E\varepsilon_{yy} &= \sigma_{yy} - \nu(\sigma_{xx} + \sigma_{zz}) \\
&= \sigma_{yy} - \nu(\sigma_{xx} + \nu(\sigma_{xx} + \sigma_{yy})) \\
&= -\nu(1+\nu)\sigma_{xx} + (1-\nu^2)\sigma_{yy} \\
E\varepsilon_{xy} &= (1+\nu)\sigma_{xy}
\end{aligned}
$$

or,

$$
\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} = \frac{1}{E} \begin{pmatrix} 1-\nu^2 & -\nu(1+\nu) & 0 \\ -\nu(1+\nu) & 1-\nu^2 & 0 \\ 0 & 0 & 1+\nu \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} \tag{22.12}
$$

or[4]

$$
\begin{aligned}
\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} &= \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 1-2\nu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \\
&= \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & \frac{1}{2}(1-2\nu) \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ 2\varepsilon_{xy} \end{pmatrix} \tag{22.13}
\end{aligned}
$$

We then have

$$
\boldsymbol{D}_{\text{plane strain}} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & \frac{1}{2}(1-2\nu) \end{pmatrix} \tag{22.14}
$$

Remark. *The compliance matrix for plane strain is not found by removing columns and rows from the general isotropic compliance matrix!*

Let us also look at another notation used in Simpson's book [1250] (Eq. 12.3). We start from Eq. (22.3):

---

[4]https://www.efunda.com/formulae/solid_mechanics/mat_mechanics/hooke_plane_strain.cfm

$$\begin{aligned}
\sigma_{xx} &= \lambda(\varepsilon_{xx} + \varepsilon_{yy}) + 2\mu\varepsilon_{xx} \\
&= (\lambda + 2\mu)\varepsilon_{xx} + \lambda\varepsilon_{yy} \\
\sigma_{yy} &= \lambda(\varepsilon_{xx} + \varepsilon_{yy}) + 2\mu\varepsilon_{yy} \\
&= \lambda\varepsilon_{xx} + (\lambda + 2\mu)\varepsilon_{yy} \\
\sigma_{xy} &= 2\mu\varepsilon_{xy}
\end{aligned} \tag{22.15}$$

so that

$$\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} = \begin{pmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ 2\varepsilon_{xy} \end{pmatrix}$$

Since $K = \lambda + 2\mu/3$, we also have

$$\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} = \begin{pmatrix} K + \frac{4}{3}\mu & K - \frac{2}{3}\mu & 0 \\ K - \frac{2}{3}\mu & K + \frac{4}{3}\mu & 0 \\ 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ 2\varepsilon_{xy} \end{pmatrix}$$

or,

$$\boldsymbol{D}_{\text{plane strain}} = \begin{pmatrix} K + \frac{4}{3}\mu & K - \frac{2}{3}\mu & 0 \\ K - \frac{2}{3}\mu & K + \frac{4}{3}\mu & 0 \\ 0 & 0 & \mu \end{pmatrix}$$

The stress tensor is then as follows:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & 0 \\ \sigma_{yx} & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{pmatrix}$$

and its second moment invariant is given by

$$\mathcal{I}_2(\boldsymbol{\sigma}) = \frac{1}{2}\boldsymbol{\sigma} : \boldsymbol{\sigma} = \frac{1}{2}(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2) + \sigma_{xy}^2$$

The deviatoric stress is given by $\boldsymbol{\tau} = \boldsymbol{\sigma} - \frac{1}{3}\text{tr}(\boldsymbol{\sigma})\boldsymbol{1}$ with in this case

$$\begin{aligned}
\text{tr}(\boldsymbol{\sigma}) &= \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \\
&= \sigma_{xx} + \sigma_{yy} + \nu(\sigma_{xx} + \sigma_{yy}) \\
&= (1 + \nu)(\sigma_{xx} + \sigma_{yy})
\end{aligned} \tag{22.16}$$

so that

$$\begin{aligned}
\boldsymbol{\tau} &= \boldsymbol{\sigma} - \frac{1+\nu}{3}(\sigma_{xx} + \sigma_{yy})\boldsymbol{1} \\
&= \frac{1}{3}\begin{pmatrix} 3\sigma_{xx} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 3\sigma_{xy} & 0 \\ 3\sigma_{yx} & 3\sigma_{yy} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 0 \\ 0 & 0 & 3\sigma_{zz} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) \end{pmatrix} \\
&= \frac{1}{3}\begin{pmatrix} 3\sigma_{xx} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 3\sigma_{xy} & 0 \\ 3\sigma_{yx} & 3\sigma_{yy} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 0 \\ 0 & 0 & 3\nu(\sigma_{xx} + \sigma_{yy}) - (1+\nu)(\sigma_{xx} + \sigma_{yy}) \end{pmatrix} \\
&= \frac{1}{3}\begin{pmatrix} 3\sigma_{xx} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 3\sigma_{xy} & 0 \\ 3\sigma_{yx} & 3\sigma_{yy} - (1+\nu)(\sigma_{xx} + \sigma_{yy}) & 0 \\ 0 & 0 & (2\nu - 1)(\sigma_{xx} + \sigma_{yy}) \end{pmatrix}
\end{aligned} \tag{22.17}$$

and also

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{2}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2) + \tau_{xy}^2$$

1158

Remark. *In the case of a (near-)incompressible material, $\nu \to \frac{1}{2}$ then $\tau_{zz} \to 0$.*

22.3   Plane stress

For thin geometries. Let z be the direction perpendicular to the plate. Tractions on the z-surface are assumed to be negligible, e.g. $\sigma_{zz} = \sigma_{yz} = \sigma_{xz} = 0$

$$
\begin{aligned}
E\varepsilon_{xx} &= \sigma_{xx} - \nu(\sigma_{yy} + \sigma_{zz}) \\
E\varepsilon_{yy} &= \sigma_{yy} - \nu(\sigma_{xx} + \sigma_{zz}) \\
E\varepsilon_{zz} &= \sigma_{zz} - \nu(\sigma_{xx} + \sigma_{yy}) \\
E\varepsilon_{xy} &= (1+\nu)\sigma_{xy} \\
E\varepsilon_{xz} &= (1+\nu)\sigma_{xz} \\
E\varepsilon_{yz} &= (1+\nu)\sigma_{yz}
\end{aligned}
$$

Immediately we have $\varepsilon_{xz} = \varepsilon_{yz} = 0$. Furthermore,

$$
E\varepsilon_{xx} + E\varepsilon_{yy} = \sigma_{xx} - \nu\sigma_{yy} + \sigma_{yy} - \nu\sigma_{xx} = (1-\nu)(\sigma_{xx} + \sigma_{yy})
$$

so that the third equation can be written

$$
\varepsilon_{zz} = -\frac{\nu}{1-\nu}(\varepsilon_{xx} + \varepsilon_{yy})
$$

Then,

$$
\begin{aligned}
E\varepsilon_{xx} &= \sigma_{xx} - \nu\sigma_{yy} \\
E\varepsilon_{yy} &= \sigma_{yy} - \nu\sigma_{xx} \\
E\varepsilon_{xy} &= (1+\nu)\sigma_{xy}
\end{aligned}
\tag{22.18}
$$

or,

$$
\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} = \frac{1}{E} \begin{pmatrix} 1 & -\nu & 0 \\ -\nu & 1 & 0 \\ 0 & 0 & 1+\nu \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix}
\tag{22.19}
$$

or[5],

$$
\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} = \frac{E}{(1-\nu^2)} \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 1-\nu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}
\tag{22.20}
$$

We then have

$$
\boldsymbol{D}_{\text{plane stress}} = \frac{E}{(1-\nu^2)} \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1}{2}(1-\nu) \end{pmatrix}
\tag{22.21}
$$

Remark. *The stiffness matrix for plane stress is not found by removing columns and rows from the general isotropic stiffness matrix.*

---

[5]https://www.efunda.com/formulae/solid_mechanics/mat_mechanics/hooke_plane_stress.cfm

22.4 The axisymmetric case

We start from

$$\boldsymbol{\sigma} = \lambda(\vec{\nabla} \cdot \vec{v})\,\mathbf{1} + 2\mu\boldsymbol{\varepsilon}(\vec{v}) \tag{22.22}$$

In cylindrical coordinates the velocity gradient is given by

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & \frac{1}{r}\frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r} & \frac{\partial v_r}{\partial z} \\ \\ \frac{\partial v_\theta}{\partial r} & \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r} & \frac{\partial v_\theta}{\partial z} \\ \\ \frac{\partial v_z}{\partial r} & \frac{1}{r}\frac{\partial v_z}{\partial \theta} & \frac{\partial v_z}{\partial z} \end{pmatrix}$$

In the case of axisymmetry, and in this case symmetry about the z axis, there is invariance with respect to the rotation around the axis so stresses and other quantities are independent of the θ coordinate, or simply put $\partial_\theta \to 0$. The velocity gradient simplifies to:

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & -\frac{v_\theta}{r} & \frac{\partial v_r}{\partial z} \\ \\ \frac{\partial v_\theta}{\partial r} & \frac{v_r}{r} & \frac{\partial v_\theta}{\partial z} \\ \\ \frac{\partial v_z}{\partial r} & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix}$$

Also, it follows logically that $v_\theta = 0$ so that ultimately:

$$\vec{\nabla}\vec{v} = \begin{pmatrix} \frac{\partial v_r}{\partial r} & 0 & \frac{\partial v_r}{\partial z} \\ \\ 0 & \frac{v_r}{r} & 0 \\ \\ \frac{\partial v_z}{\partial r} & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix}$$

and the strain tensor is then given by

$$\boldsymbol{\varepsilon}(\vec{v}) = \frac{1}{2}\left(\vec{\nabla}\vec{v} + \vec{\nabla}\vec{v}^T\right) = \begin{pmatrix} \frac{\partial v_r}{\partial r} & 0 & \frac{1}{2}\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right) \\ \\ 0 & \frac{v_r}{r} & 0 \\ \\ \frac{1}{2}\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right) & 0 & \frac{\partial v_z}{\partial z} \end{pmatrix} \tag{22.23}$$

The term $\vec{\nabla} \cdot \vec{v}$ is simply the trace of $\boldsymbol{\varepsilon}(\vec{v})$ so

$$\vec{\nabla} \cdot \vec{v} = \frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z}$$

Finally the full stress tensor is then

$$\boldsymbol{\sigma} = \begin{pmatrix} \lambda(\frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z}) + 2\mu\frac{\partial v_r}{\partial r} & 0 & \mu(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}) \\ \\ 0 & \lambda(\frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z}) + 2\mu\frac{v_r}{r} & 0 \\ \\ \mu(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}) & 0 & \lambda(\frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z}) + 2\mu\frac{\partial v_z}{\partial z} \end{pmatrix}$$

$$= \begin{pmatrix} (\lambda + 2\mu)\frac{\partial v_r}{\partial r} + \lambda(\frac{v_r}{r} + \frac{\partial v_z}{\partial z}) & 0 & \mu(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}) \\ \\ 0 & (\lambda + 2\mu)\frac{v_r}{r} + \lambda(\frac{\partial v_r}{\partial r} + \frac{\partial v_z}{\partial z}) & 0 \\ \\ \mu(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}) & 0 & (\lambda + 2\mu)\frac{\partial v_z}{\partial z} + \lambda(\frac{\partial v_r}{\partial r} + \frac{v_r}{r}) \end{pmatrix}$$

As we did in the 2D case, we rewrite the six independent stress terms in to a vector $\vec{\sigma}$ and we use Eq. (22.22) to arrive at:

$$\vec{\sigma} = \begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{r\theta} \\ \sigma_{rz} \\ \sigma_{\theta z} \end{pmatrix} = \begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{r\theta} \\ 2\varepsilon_{rz} \\ 2\varepsilon_{\theta z} \end{pmatrix} = \vec{\varepsilon}(\vec{\upsilon})$$

or $\vec{\sigma} = \boldsymbol{D} \cdot \vec{\varepsilon}(\vec{\upsilon})$. The components of the $\vec{\varepsilon}(\vec{\upsilon})$ vector are

$$\vec{\varepsilon}(\vec{\upsilon}) = \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{r\theta} \\ 2\varepsilon_{rz} \\ 2\varepsilon_{\theta z} \end{pmatrix} = \begin{pmatrix} \frac{\partial \upsilon_r}{\partial r} \\ \frac{\upsilon_r}{r} \\ \frac{\partial \upsilon_z}{\partial z} \\ 0 \\ \frac{\partial \upsilon_z}{\partial r} + \frac{\partial \upsilon_r}{\partial z} \\ 0 \end{pmatrix}$$

We see that there are two zeroes and consequently we'll find that $\sigma_{r\theta}$ and $\sigma_{\theta z}$ are also identically zero, so we discard these and end up with only four stress components :

$$\vec{\sigma} = \begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix} = \begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{rz} \end{pmatrix}$$

Note that in the literature the above relationship is often written

$$\begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \lambda & \nu & 0 \\ \nu & 1-\nu & \nu & 0 \\ \nu & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & (1-2\nu)/2 \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{rz} \end{pmatrix}$$

which is equivalent since $E = 2\mu(1+\nu)$ and $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}$ (see for instance Section 5.2.4 in [1540]).

---

about the implementation:

Only displacements in the r and z directions remain (note that $\varepsilon_{\theta\theta}$ is in fact equal to υ_r/r). In what follows I rename $u = \upsilon_r$ and $\upsilon_z = w$ to simplify notations. Then, inside an element we have

$$u^h(r,z) = \sum_{i=1}^{m} \mathcal{N}_i(r,z) u_i$$

$$w^h(r,z) = \sum_{i=1}^{m} \mathcal{N}_i(r,z) w_i \tag{22.24}$$

where $\mathcal{N}_i$ are the basis functions attached to the m nodes of the element. We compute the elements

of the $\boldsymbol{\varepsilon}$ tensor of Eq. (22.23) as follows:

$$\varepsilon_{rr} = \frac{\partial u^h}{\partial r} = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r, z)\, u_i \tag{22.25}$$

$$\varepsilon_{\theta\theta} = \frac{u_r^h}{r} = \frac{1}{r}\sum_{i=1}^{m} \mathcal{N}_i(r, z)\, u_i \tag{22.26}$$

$$\varepsilon_{zz} = \frac{\partial w^h}{\partial z} = \sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial z}(r, z)\, w_i \tag{22.27}$$

$$\varepsilon_{rz} = \frac{1}{2}\frac{\partial u^h}{\partial z} + \frac{1}{2}\frac{\partial w^h}{\partial r} = \frac{1}{2}\sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial z}(r, z)u_i + \frac{1}{2}\sum_{i=1}^{m} \frac{\partial \mathcal{N}_i}{\partial r}(r, z)w_i \tag{22.28}$$

Let us take $m = 3$, i.e. linear triangles, for simplicity. Then the strain vector $\vec{\varepsilon}^h$ is given by

$$\vec{\varepsilon}^h = \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\theta\theta} \\ \varepsilon_{zz} \\ 2\varepsilon_{rz} \end{pmatrix} = \begin{pmatrix} \frac{\partial u^h}{\partial r} \\ \frac{u^h}{r} \\ \frac{\partial w^h}{\partial z} \\ \frac{\partial u^h}{\partial z} + \frac{\partial w^h}{\partial r} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial \mathcal{N}_1}{\partial r} & 0 & \frac{\partial \mathcal{N}_2}{\partial r} & 0 & \frac{\partial \mathcal{N}_3}{\partial r} & 0 \\ \frac{\mathcal{N}_1}{r} & 0 & \frac{\mathcal{N}_2}{r} & 0 & \frac{\mathcal{N}_3}{r} & 0 \\ 0 & \frac{\partial \mathcal{N}_1}{\partial z} & 0 & \frac{\partial \mathcal{N}_2}{\partial z} & 0 & \frac{\partial \mathcal{N}_3}{\partial z} \\ \frac{\partial \mathcal{N}_1}{\partial z} & \frac{\partial \mathcal{N}_1}{\partial r} & \frac{\partial \mathcal{N}_2}{\partial z} & \frac{\partial \mathcal{N}_2}{\partial r} & \frac{\partial \mathcal{N}_3}{\partial z} & \frac{\partial \mathcal{N}_3}{\partial r} \end{pmatrix}}_{\boldsymbol{B}(4\times6)} \cdot \underbrace{\begin{pmatrix} u1 \\ w1 \\ u2 \\ w2 \\ u3 \\ w3 \end{pmatrix}}_{\vec{\mathcal{U}}(6\times1)}$$

or $\vec{\varepsilon}^h = \boldsymbol{B} \cdot \vec{\mathcal{U}}$ and finally

$$\underbrace{\begin{pmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{pmatrix}}_{\vec{\sigma}} = \underbrace{\begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix}}_{\boldsymbol{D}} \cdot \underbrace{\begin{pmatrix} \frac{\partial \mathcal{N}_1}{\partial r} & 0 & \frac{\partial \mathcal{N}_2}{\partial r} & 0 & \frac{\partial \mathcal{N}_3}{\partial r} & 0 \\ \frac{\mathcal{N}_1}{r} & 0 & \frac{\mathcal{N}_2}{r} & 0 & \frac{\mathcal{N}_3}{r} & 0 \\ 0 & \frac{\partial \mathcal{N}_1}{\partial z} & 0 & \frac{\partial \mathcal{N}_2}{\partial z} & 0 & \frac{\partial \mathcal{N}_3}{\partial z} \\ \frac{\partial \mathcal{N}_1}{\partial z} & \frac{\partial \mathcal{N}_1}{\partial r} & \frac{\partial \mathcal{N}_2}{\partial z} & \frac{\partial \mathcal{N}_2}{\partial r} & \frac{\partial \mathcal{N}_3}{\partial z} & \frac{\partial \mathcal{N}_3}{\partial r} \end{pmatrix}}_{\boldsymbol{B}(4\times6)} \cdot \underbrace{\begin{pmatrix} u1 \\ w1 \\ u2 \\ w2 \\ u3 \\ w3 \end{pmatrix}}_{\vec{\mathcal{U}}(6\times1)}$$

or,

$$\boxed{\vec{\sigma} = \boldsymbol{D} \cdot \boldsymbol{B} \cdot \vec{\mathcal{U}}}$$

Note that in 2D, the matrix $\boldsymbol{D}$ is 3×3 and $\boldsymbol{B}$ is 3×6.

`I do not know yet how to arrive at what follows`

The 6×6 stiffness matrix is then

$$\mathbb{K} = \iiint \boldsymbol{B}^T \cdot \boldsymbol{D} \cdot \boldsymbol{B}\, dV$$

with $dV = rdrd\theta dz$ in cylindrical coordinates. The integral over the θ coordinate yields a factor 2π so

$$\mathbb{K} = 2\pi \iint \boldsymbol{B}^T \cdot \boldsymbol{D} \cdot \boldsymbol{B}\, rdrdz$$

The integration can now be performed as simply as was the case in the plane stress problem.

`write the derivation for the rhs`

Note that in practice the matrix $\boldsymbol{D}$ is computed as follows (see for example Stone 63):

$$\boldsymbol{D} = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix} = \lambda \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \mu \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

22.5 FEM: Incompressible formulation from Zienkiewicz & Taylor book

This is from Volume 1-The Basis, page 307. Note that the authors use a different sign convention for pressure so that what follows is adapted from the book.

The authors start by defining the vector $\vec{m}^T = (1, 1, 1, 0, 0, 0)$. Using the vector notation of stress, the mean stress or pressure is given by

$$p = -\frac{1}{3}\text{tr}(\boldsymbol{\sigma}) = -\frac{1}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) = -\frac{1}{3}(1, 1, 1, 0, 0, 0) \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} = -\frac{1}{3}\vec{m}^T \cdot \vec{\sigma}$$

For isotropic behaviour the 'pressure' is related to the volumetric strain by the bulk modulus of the material, K. Thus,

$$p = -K\text{tr}(\boldsymbol{\varepsilon}(\vec{v})) = -K\vec{m}^T \cdot \vec{\varepsilon}(\vec{v})$$

For an incompressible material $K \to \infty$ and the volumetric strain is simply zero. Since $\text{tr}(\boldsymbol{\varepsilon}(\vec{v})) = \vec{m}^T \cdot \vec{\varepsilon}(\vec{v})$ the deviatoric strain is defined by

$$\vec{\varepsilon}^d(\vec{v}) = \vec{\varepsilon}(\vec{v}) - \frac{1}{3}\vec{m}\,\text{tr}(\boldsymbol{\varepsilon}(\vec{v})) = \vec{\varepsilon}(\vec{v}) - \frac{1}{3}\vec{m}\vec{m}^T \cdot \vec{\varepsilon}(\vec{v}) = \left(\mathbf{1} - \frac{1}{3}\vec{m}\vec{m}^T\right) \cdot \vec{\varepsilon}(\vec{v}) = \boldsymbol{I}_d \cdot \vec{\varepsilon}(\vec{v})$$

where $\boldsymbol{I}_d$ is the already defined deviatoric projection matrix (see Eq. (22.10)).

change notation to big lambda matrix?

In isotropic elasticity the deviatoric strain is related to the deviatoric stress by the shear modulus

μ as $\boldsymbol{\tau} = 2\mu\boldsymbol{\varepsilon}^d(\vec{v})$, or

$$
\vec{\tau} = \underbrace{\begin{pmatrix} 2\mu & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}}_{\boldsymbol{C}_\mu} \cdot \vec{\varepsilon}^{\,d}(\vec{v})
$$

$$
= \boldsymbol{C}_\mu \cdot \boldsymbol{I}_d \cdot \vec{\varepsilon}(\vec{v})
$$

$$
= \boldsymbol{C}_\mu \cdot \left(\mathbf{1} - \frac{1}{3}\vec{m}\vec{m}^T\right) \cdot \vec{\varepsilon}(\vec{v})
$$

$$
= \boldsymbol{C}_\mu \cdot \vec{\varepsilon}(\vec{v}) - \frac{1}{3} \begin{pmatrix} 2\mu & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \vec{\varepsilon}(\vec{v})
$$

$$
= \boldsymbol{C}_\mu \cdot \vec{\varepsilon}(\vec{v}) - \frac{1}{3} \begin{pmatrix} 2\mu & 2\mu & 2\mu & 0 & 0 & 0 \\ 2\mu & 2\mu & 2\mu & 0 & 0 & 0 \\ 2\mu & 2\mu & 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \vec{\varepsilon}(\vec{v})
$$

$$
= \underbrace{\left(\boldsymbol{C}_\mu - \frac{2\mu}{3}\vec{m}\vec{m}^T\right)}_{\boldsymbol{D}_d} \cdot \vec{\varepsilon}(\vec{v}) \tag{22.29}
$$

We can also write $\boldsymbol{D}_d$ in a more explicit manner more amenable to implementation:

$$
\boldsymbol{D}_d = \frac{\mu}{3} \begin{pmatrix} 4 & -2 & -2 & 0 & 0 & 0 \\ -2 & 4 & -2 & 0 & 0 & 0 \\ -2 & -2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix} \tag{22.30}
$$

> CHANGE matrix names/notations!

In the mixed form considered next we shall use as variables the displacement $\vec{v}$ and the pressure p. Following the usual approach we arrive at a linear system $\mathcal{A} \cdot \vec{X} = \vec{b}$ with (may be be a bit careful

about signs again)

$$\vec{X} = \begin{pmatrix} \vec{\mathcal{U}} \\ \vec{\mathcal{P}} \end{pmatrix}$$

$$\vec{b} = \begin{pmatrix} \vec{f} \\ \vec{0} \end{pmatrix}$$

$$\mathcal{A} = \begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & -\mathbb{M}_K \end{pmatrix}$$

$$\mathbb{K} = \int_\Omega \boldsymbol{B}^T \cdot \boldsymbol{D}_d \cdot \boldsymbol{B} \, dV$$

$$\mathbb{G} = \int_\Omega \boldsymbol{B}^T \boldsymbol{N}_p \, dV$$

$$\mathbb{M}_K = \int_\Omega \frac{1}{K} \boldsymbol{N}_p^T \cdot \boldsymbol{N}_p \, dV$$

Remark. *A similar approach is taken in Sanan, May, Bollhöfer, and Schenk [1182] (2020). However the authors define a new auxiliary pressure $p = -\lambda \vec{\nabla} \cdot \vec{\mathfrak{v}}$ (as opposed to $p = -K \vec{\nabla} \cdot \vec{\mathfrak{v}}$ above).*

22.6 Elastic parameter values for Earth materials

what are E, ν, μ, etc ... for Earth ? bounds ? etc ...

material	Young's modulus (in 10^6 bars)	shear modulus (in 10^6 bars)
ice	0.1	0.03
shale	0.2–0.3	0.15
limestone	0.4–0.7	0.22–0.26
granite	0.3–0.6	0.2
basalt	0.7–0.9	0.3
steel	2.1	0.83

Taken from `https://www.britannica.com/science/rock-geology/Stress-strain-relationships`

Note that $1\text{bar} = 0.1\text{MPa}$, so $10^6\text{bar}=10^5\text{MPa}$.

In Farrington, Moresi, and Capitanio [412] (2014), we find "Elastic properties within the lithosphere and mantle are relatively well constrained with a shear modulus between 10^{10} and 10^{11} Pa."

22.7 Benchmarks and analytical solutions

- Cook's membrane problem (see Lamichhane [786] (2009) and refs therein; see Lamichhane [785] (2014)).

- Rectangular beam problem (see Lamichhane [786] (2009) and refs therein; see Lamichhane [785] (2014)).

- Thick-walled sphere under internal pressure (see Lamichhane [786] (2009) and refs therein).

22.7.1 Cantilever beam

To be found in Zhang, Wang, and Chan [1515] (2014). A cantilever beam subjected to a parabolic traction $P = 1$ at the free end is considered as shown in the following figure. The geometry is $L = 10$, $D = 2$.
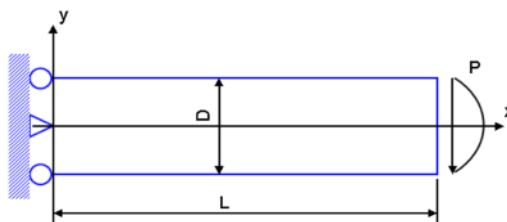
The exact displacement solution of the problem is given by

$$
\begin{aligned}
u_x &= \frac{Py}{6\bar{E}I}\left[(6L-3x)x + (2+\bar{\nu})(y^2 - D^2/4)\right]\\
u_y &= -\frac{P}{\bar{E}I}\left[3\bar{\nu}y^2(L-x) + (4+5\bar{\nu})\frac{D^2x}{4} + (3L-x)x^2\right]
\end{aligned}
\tag{22.31}
$$

where

- plane stress: $\bar{E} = E$, $\bar{\nu} = \nu$

- plane strain: $\bar{E} = E/(1-\nu^2)$, $\bar{\nu} = \nu(1-\nu)$

The plane strain condition is studied for volumetric locking, and $I = D^3/12$ is the second moment of area of the beam. The corresponding stress solutions are

$$
\begin{aligned}
\sigma_{xx}(x, y) &= \frac{P(L - x)y}{I} \\
\sigma_{yy}(x, y) &= 0 \\
\sigma_{xy}(x, y) &= -\frac{P}{2I}\left(D^2/4 - y^2\right)
\end{aligned}
\tag{22.32}
$$

In computation, analytical displacement solution obtained from Equation (22.31) is prescribed along the left boundary, and the parabolic traction on the right boundary applies the exact shear stress solution provided in Equation (22.32). In the paper the authors use a coarse discretization of 2×10 elements.

22.7.2 Infinite plate with a circular hole

22.7.3 Cook's membrane

22.7.4 Punching problem

Chapter 23

[WIP] Heat Transfer & convection in a porous medium

This chapter is currently being developed outside of FieldStone.

Chapter 24

[WIP] Adjoint methods

adjoint methods in geodynamics [202, 486, 623, 658, 826, 1440, 1437, 1469]. Also see `johnson_Notes on Adjoint Methods.pdf` and `bradley-PDE-constrained optimization and the adjoint method.pdf`

Advection-diffusion: `https://en.wikipedia.org/wiki/Adjoint_equation`

Derivation of the adjoint poisson equation: `https://math.stackexchange.com/questions/2269111/derivation-of-the-adjoint-poisson-equation`

Inverting PDEs with adjoints (esp Poisson) `https://joelcfd.com/inverting-pdes-with-adjoints/`

Video: Introduction to the adjoint method `https://youtu.be/EybH_Q-QTZ8`

Video: adjoint-based optimization `https://youtu.be/Yiz92Ekn7vU`

Chapter 25

Visco-elasticity: physics, formulations and FEM

Work in progress!!!

25.1   A remark

In the absence of gravity, density does not enter the equations. As such compressible elasticity is fine.

If gravitational forces are taken into account, compressible elasticity means that density should change in space/time. However, a lot of the experiments hereafter are written with a constant density, meaning that they are reproducible or valid only for incompressible elastic (and viscou) flows!

25.2   Analytical Benchmarks

25.2.1   the 1D solution

We wish to find the general solution $\tau(t)$ of from the first order ODE

$$\frac{1}{2\mu}\frac{d\tau}{dt} + \frac{1}{2\eta}\tau = \dot{\varepsilon}_0 \qquad \text{or,} \qquad \frac{d\tau}{dt} + \frac{\mu}{\eta}\tau = 2\mu\dot{\varepsilon}_0$$

There is a standard technique to solve such equations, and we start with the following equation instead:

$$\frac{1}{2\mu}\frac{d\tau}{dt} + \frac{1}{2\eta}\tau = 0$$

It can be rewritten

$$\frac{d\tau}{\tau} = -\frac{\mu}{\eta}dt$$

so

$$\int \frac{d\tau}{\tau} = -\frac{\mu}{\eta}\int dt$$

$$\ln\tau - \ln\tau_0 = -\frac{\mu}{\eta}t$$

$$\ln\tau = -\frac{\mu}{\eta}t + \ln\tau_0$$

$$\tau(t) = \exp\left(-\frac{\mu}{\eta}t + \ln\tau_0\right) = \exp\left(-\frac{\mu}{\eta}t\right)\exp\left(\ln\tau_0\right) = \tau_0\exp\left(-\frac{\mu}{\eta}t\right)$$

or, using the Maxwell time $t_M = \eta/\mu$,

$$\tau(t) = \tau_0 \exp\left(-\frac{t}{t_M}\right)$$

Based on this solution we now consider the following equation

$$\frac{d}{dt}\left[\tau \exp\left(\frac{t}{t_M}\right)\right] = \frac{d\tau}{dt}\exp\left(\frac{t}{t_M}\right) + \tau\frac{1}{t_M}\exp\left(\frac{t}{t_M}\right)$$

$$= \left(\underbrace{\frac{d\tau}{dt} + \frac{\mu}{\tau}\tau}_{2\mu\dot{\varepsilon}_0}\right)\exp\left(\frac{t}{t_M}\right)$$

Then we proceed to integrate both sides:

$$\int \frac{d}{dt}\left[\tau \exp\left(\frac{t}{t_M}\right)\right] dt = 2\mu\dot{\varepsilon}_0 \int \exp\left(\frac{t}{t_M}\right) dt$$

$$\tau(t)\exp\left(\frac{t}{t_M}\right) - \tau(0)\exp\left(\frac{0}{t_M}\right) = 2\mu\dot{\varepsilon}_0 t_M[\exp\left(\frac{t}{t_M}\right) - \exp\left(\frac{0}{t_M}\right)]$$

$$\tau(t)\exp\left(\frac{t}{t_M}\right) - \tau(0) = 2\mu\dot{\varepsilon}_0\frac{\eta}{\mu}[\exp\left(\frac{t}{t_M}\right) - 1]$$

$$\tau(t)\exp\left(\frac{t}{t_M}\right) = 2\mu\dot{\varepsilon}_0\frac{\eta}{\mu}[\exp\left(\frac{t}{t_M}\right) - 1] + \tau(0)$$

$$\tau(t) = 2\eta\dot{\varepsilon}_0[1 - \exp\left(-\frac{t}{t_M}\right)] + \tau(0)\exp\left(-\frac{t}{t_M}\right)$$

which is the solution of Eq. (4) of Kaus and Becker [721] (2007).

Alternative:
The general solution can be arrived at by means of the Laplace transform (?!) and is given by:

$$\boldsymbol{\tau}(t) = \boldsymbol{\tau}(t_0)\exp\left(-\frac{t - t_0}{t_M}\right) + \exp\left(-\frac{t}{t_M}\right)\int_{t_0}^t 2\mu\dot{\boldsymbol{\varepsilon}}_T \exp\left(\frac{t'}{t_M}\right) dt'$$

If $t_0 =$ and $\boldsymbol{\tau}(t_0) = 0$ then

$$\boldsymbol{\tau}(t) = \exp\left(-\frac{t}{t_M}\right)\int_0^t 2\mu\dot{\boldsymbol{\varepsilon}}_T \exp\left(\frac{t'}{t_M}\right) dt'$$

If the strain rate and shear modulus are constant in time, then

$$\boldsymbol{\tau}(t) = \exp\left(-\frac{t}{t_M}\right) 2\mu\dot{\boldsymbol{\varepsilon}}_T \int_0^t \exp\left(\frac{t'}{t_M}\right) dt'$$

$$= \exp\left(-\frac{t}{t_M}\right) 2\mu\dot{\boldsymbol{\varepsilon}}_T t_M \left[\exp\left(\frac{t}{t_M}\right) - 1\right]$$

$$= 2\eta\dot{\boldsymbol{\varepsilon}}_T \left[1 - \exp\left(-\frac{t}{t_M}\right)\right]$$

since $t_M = \eta/\mu$.

25.2.2 Pure shear

The first benchmark performed to test the viscoelastic implementation considers the stress build-up present in a viscoelastic Maxwell body. Contrary to stressed viscous materials, viscoelastic materials gradually build-up stress when sheared after which a transition to viscous deformation occurs.

An unstressed, incompressible viscoelastic Maxwell medium is subjected to a velocity field resulting in pure shear. The increase of the accumulated stress with time is given by an analytical solution:

$$\boldsymbol{\tau} = 2\eta \, \dot{\boldsymbol{\varepsilon}} \left(1 - e^{-\frac{\mu}{\eta}t}\right) \tag{25.1}$$

with t time, η the prescribed material viscosity and μ the prescribed material shear modulus. The domain size is 100×100km. The velocity prescribed at all boundaries equals $v = 1$ cm/yr in magnitude yielding a constant background strain rate of $\dot{\varepsilon} = 2$cm/yr/100km $\simeq 6.342 \times 10^{-15}\,\mathrm{s}^{-1}$. The viscosity is $\eta = 10^{21}$Pa.s, the shear modulus is $\mu = 10^{10}$Pa and the gravity is set to zero. We set $\delta t = 100$yr.
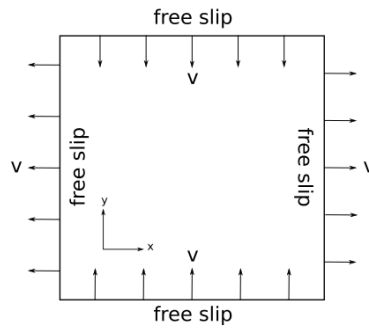
Set up of the stress build-up benchmark. All domain sides have a free slip boundary condition, and pure shear velocity conditions are prescribed. Adapted from Gerya (2010) [483].

We have

$$\eta_{eff} = \frac{\eta \delta t}{\delta t + \eta/\mu} = \frac{10^{21} \cdot 3.154 \times 10^9}{3.154 \times 10^9 + 10^{21}/10^{10}} \simeq 3.0592 \times 10^{19}\mathrm{Pa.s} \qquad \text{and} \qquad Z = \frac{\eta_{eff}}{\mu \delta t} \simeq 0.9694$$

The Maxwell time is $t_M = \frac{\eta}{\mu} = 10^{11}$s $\simeq 3171$yr. In the absence of elasticity (purely viscous behaviour), we have $\dot{\varepsilon}_{xx} = 6.342 \times 10^{-15}$ and $\eta = 10^{21}$ so the deviatoric stress τ_{xx} is equal to

$$\tau_{xx} = 2 \cdot 10^{21} \cdot 6.342 \times 10^{-15} \simeq 12.68 \times 10^6 \mathrm{Pa}$$

The first time that the Stokes system is solved, there is no stored stress, i.e. the elastic rhs is identically zero, so that the system is solved with a viscosity equal to η_{eff}. We can easily compute the analytical solution, and we see that $\dot{\varepsilon}_{xy} = 0$ and $\dot{\omega}_{xy} = 0$, which we recover:

Results in Stone 64!

The expected stress value for τ_{xx} after the first Stokes solve is

$$\tau_{xx} = 2\eta_{eff}\dot{\varepsilon}_{xx} = 2 \cdot 3.057 \times 10^{19} \cdot 6.342 \times 10^{-15} \simeq 38.775 \times 10^4 \mathrm{Pa}$$

Also check Gerya's book page 358 2nd edition ?

25.2.3 simple shear

25.2.4 Rayleigh-Taylor instability

This experiment is presented in Kaus and Becker [721] (2007).

The model consists of a viscoelastic layer of thickness H_1, with density ρ_1, viscosity η_1 and elastic shear module μ_1 that overlies a viscous layer of thickness $H - H_1$, with density ρ_2 and viscosity μ_2. The interface between the two fluids is perturbed sinusoidally according to $h(x) = (H - H_1) + A_0 \cos(2\pi x/\lambda)$, where H_1 is the thickness of the upper layer, H the height of the model, A_0 the initial amplitude and λ the wavelength of the perturbation. If $\rho_1 > \rho_2$, the system is gravitationally unstable.
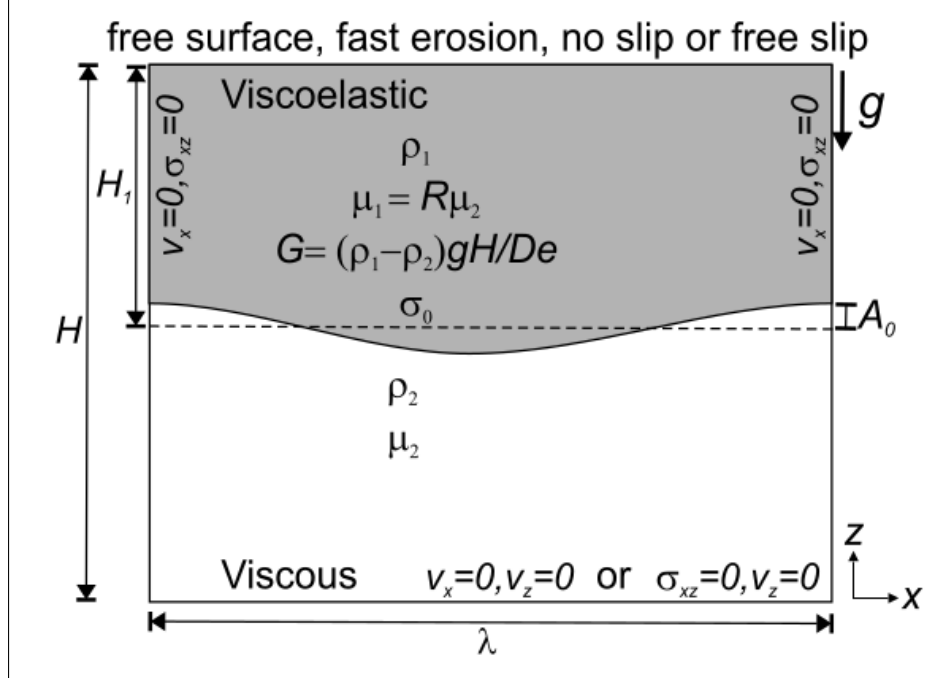
The Deborah number is given by

$$De = \frac{(\rho_1 - \rho_2)gH}{\mu}$$

and is here defined as the ratio between the viscous (Stokes) timescale $((\rho_1 - \rho_2)gH/\eta_1)$ and the viscoelastic timescale η_1/μ of the upper layer. The authors state: " Interestingly, the Deborah number, which is a measure of the importance of elasticity [...], is independent on the viscosity of the system. This is due to the fact that the magnitude of stress is solely dependent on the density difference for purely buoyancy-driven flow." Also: " In the present definition of the Deborah number realistic values for lithospheric-scale deformation are $10^{-4} \leq De \leq 1$ (with $\rho = 10 - 330$ kg m^{-3}, $g = 10$ m s^{-2}, $H = 100 - 3000$ km, $\mu = 10^{10} - 10^{11}$ Pa)."

Another parameter controlling the dynamics of the system is the viscosity contrast between the upper and the lower layer, expressed by $R = \eta_1/\eta_2$.

We here set $H = 500$ km, $H_1 = 100$ km, $\rho_2 = 3300$ kg m^{-3}, $\eta_2 = 10^{21}$ Pa s, $\mu_1 = \mu_2 = 10^{10}$ Pa, $\eta_1 = R\eta_2$, $A_0 = 1$ km, $\lambda = L_x$.

25.2.5 stress build-up inside an elastic inclusion in a viscous matrix (Beuchert & Podlachikov

fully incompressible

This is presented in Section 5.4 of Beuchert and Podladchikov [92] (2010). The authors test their viscoelastic FEM model against the analytical solution for stress built-up inside an elastic inclusion in a viscous matrix under pure shear.

1174

Note that the fluid in question is incompressible as shown by their Eq. (3).

An analytical solution applicable for the time-dependent stress evolution inside a viscoelastic inclusion embedded in a viscous matrix can be derived:

$$\tau_{xx} = -\frac{(a+1)^2 \left[-1 + \exp-\frac{2\mu at}{a^2+1}\right]}{a} \dot{\varepsilon}_{xx}^d$$

where a is the aspect ratio of the inclusion and the matrix viscosity is assumed to be unity.

For this equation to be applicable to our viscoelastic model, we prescribe a high viscosity inside the inclusion ($\eta_{inclusion} = 10^5$) and thus obtain an elastic response inside the inclusion. For the matrix, we apply a viscous rheology with viscosity $\eta_{matrix} = 1$.
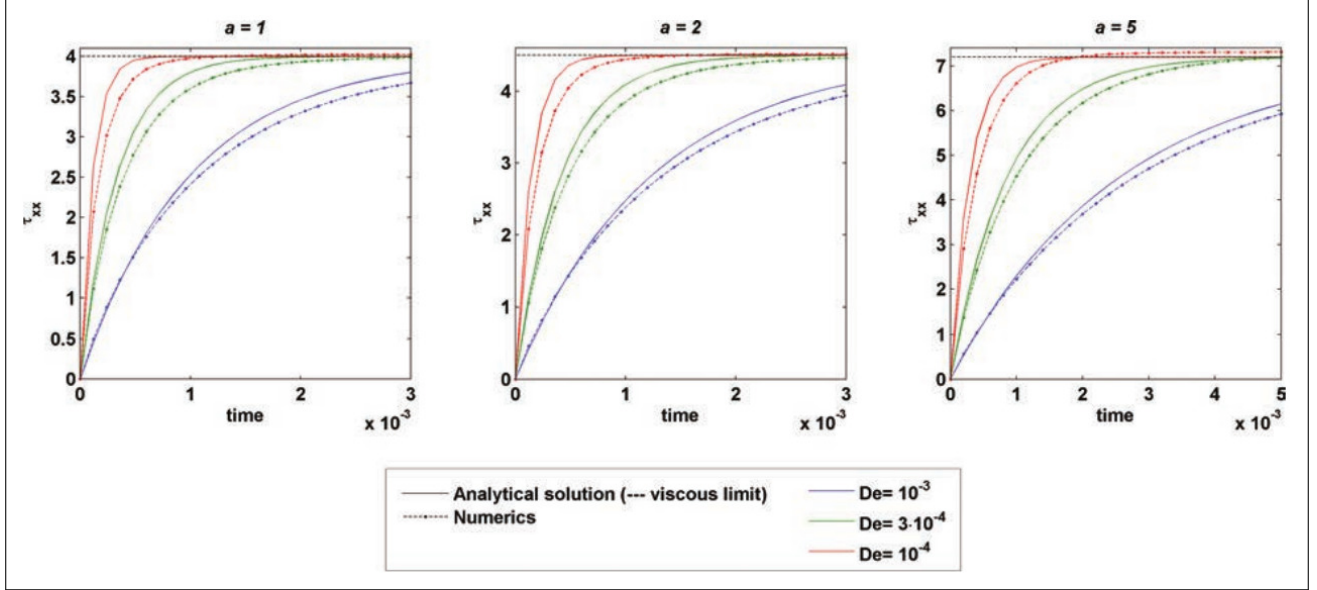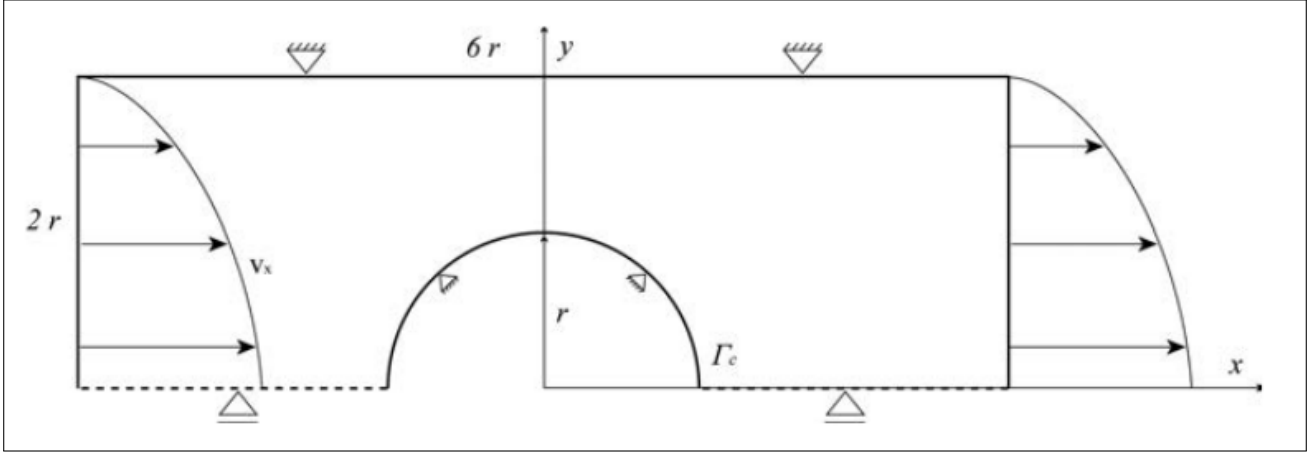

Taken from Beuchert and Podladchikov [92]. Stress built-up inside an elliptical elastic inclusion with aspect ratios $a = 1, 2, 5$ embedded in a viscous matrix under constant pure shear loading. Comparison of the analytical solution with the numerical solution obtained in their FEM model for different values of Deborah number $De = 10^{-4}, 3 \cdot 10^{-4}, 10^{-3}$ at non-dimensional deviatoric strain rate $\dot{\varepsilon}_{xx}^d = 1$. Time and deviatoric stress τ_{xx} are non-dimensional. The test results show a good agreement of the numerical solution with the analytical solution. For time $t \to \infty$, the solution for stress inside the elastic inclusion converges towards the solution for stress inside a viscous inclusion (dashed horizontal line 'viscous limit').

25.2.6 Viscoelastic flow past a cylinder in a channel (Beuchert & Podlachikov)

This is presented in Section 5.4 of Beuchert and Podladchikov [92] (2010).

We tested the flow code against a numerical benchmark for iso-viscous, viscoelastic flow past a circular cylinder in a channel. Figure below shows the domain setup and boundary conditions for this benchmark. The radius of the circular cylinder $r = 1$ is half the domain height and the domain aspect ratio is $3:1$.

At the inflow and outflow boundaries, an established Poiseuille flow $v_x = 3(R^2 - y^2)/(2R^2)$ is imposed; $v_y = 0$ at the sides. Both upper and lower boundaries are fixed in y-direction. We apply no-slip boundary conditions at the top and along the cylinder wall and free-slip (zero traction) conditions at the bottom (symmetry axis). For the inflow conditions on σ, we use the analytical solution for simple shear of a Jaumann fluid, which is valid for Poiseuille flow. In that case, the Jaumann derivative equations are

$$\tau_{xx} + 2De\omega\tau_{xy} = 0 \tag{25.2}$$
$$\tau_{yy} + 2De\omega\tau_{xy} = 0 \tag{25.3}$$
$$\tau_{xy} + 2De\omega(\tau_{yy} - \tau_{xx}) = 2\mu\dot{\varepsilon}_{xy} \tag{25.4}$$
$$\tag{25.5}$$

Given that $\dot{\gamma} = 2\dot{\varepsilon}_{xy}$ and $\omega = -1/2\dot{\gamma}$ for simple shear, we obtain from the equation above

$$\tau_{xy} = \frac{2\mu\dot{\varepsilon}_{xy}}{1 + 4De^2\omega^2} = \frac{\mu\dot{\gamma}}{1 + 4De^2\omega^2} = \frac{\mu\dot{\gamma}}{1 + 2De^2\dot{\gamma}^2}$$

and can then solve for τ_{xx} and τ_{yy} by substitution. The strain rate $\dot{\gamma}$ is given by $\partial v_x/\partial y$, that is, by differentiating the inflow condition $v_x = 3(R^2 - y^2)/(2R^2)$ with respect to y, resulting in $\dot{\gamma} = -3y/R^2$ at the inflow boundary.

The authors further measure the non-dimensional drag force C_d exerted by the passing fluid on the cylinder wall and compare it with published values.

1176
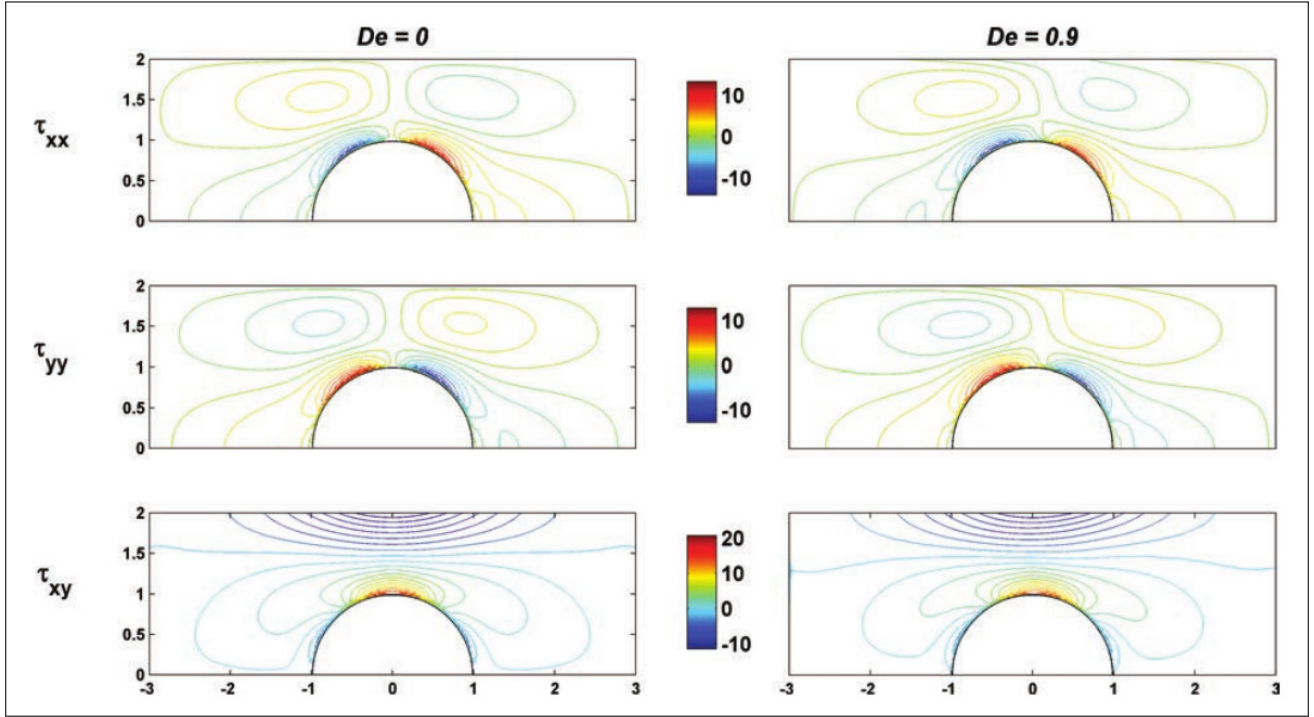
Taken from Beuchert and Podladchikov [92]. Non-dimensional deviatoric stresses τ at $De = 0$ and $De = 0.9$ resulting from fluid flow past a cylinder in a channel. The benchmark is conducted on a regular, structured grid at resolution 600×200 nodes.

25.2.7 Elastic Simple Shear, von Tscharner and Schmalholz [1428] (2015)

This experiment uses a homogeneous cube with the dimensions 1x1x1 which is deformed by simple shear. The bottom boundary is fixed (i.e., the boundary condition is no slip), the velocities on the top surface are prescribed in x-direction to generate simple shear and the top boundary is fixed in z-direction. The boundary conditions are free slip for two vertical boundaries (y=0 and y=1) and the two remaining vertical boundaries (x=0 and x=1) are free. The viscosity and the elastic shear modulus are $\eta = 10^{10}$ and $G = 1$, respectively. All model dimensions and material parameters are given in dimensionless numbers using the model length, the elastic shear modulus and the background strain rate as characteristic parameters. The results for two simulations, one with Jaumann correction and one without Jaumann correction, are given in Figure B5 where the colors indicate the second invariant of the stress tensor. The results with the Jaumann correction show a homogeneous distribution of stress (Figures a–d), whereas the distribution of the second invariant of the stress tensor is inhomogeneous without the Jaumann correction and the simulation "crashes" for high strain (Figures e–h).
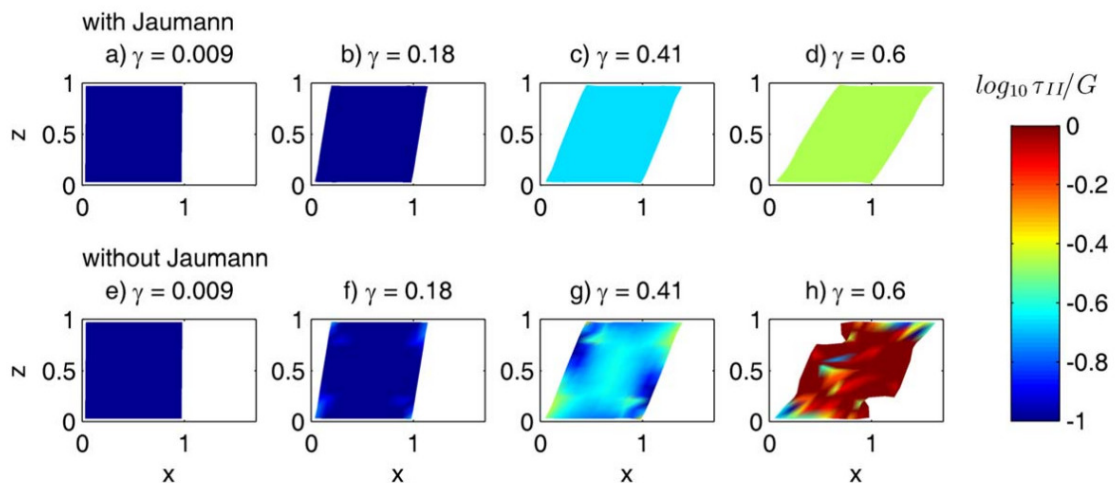
Deformation of a homogeneous cube under simple shear (a–d) with Jaumann correction and (e–h) without Jaumann correction for different amounts of bulk shear strain c (i.e., ratio of maximal horizontal displacement to model thickness). The colors indicate the second invariant of the stress tensor τ_{ii}. Without the Jaumann corrections, the stress distribution becomes inhomogeneous and the simulation crashes for high strain.

25.2.8 Response to load from ice sheet - Nakiboglu and Lambeck (1982)

The domain is 500×500km. Vertical gravity is 9.8, density is 3300, viscosity is $3 \cdot 10^{20}$ Pa s, shear modulus is 10^{10} Pa, free slip on left, right and bottom. A normal stress is imposed on the top for $0 < x < 100$km. It corresponds to an ice sheet of density $\rho_i = 900$ of 1000m height. The timestep is set to 100yr. Resolution is set to 50×50 elements. Stress/traction b.c. are explained in Section 10.15.

Analytical solution is provided in Nakiboglu and Lambeck (1982) [980]. Note however that this is a 2D setup while the original solution is for a cylindrical load and also for a semi-infinite domain.

Effective viscosity is given by

$$\eta_{eff} = \frac{\eta \delta t}{\delta t + \eta/\mu} = 2.85362675546547e + 19$$
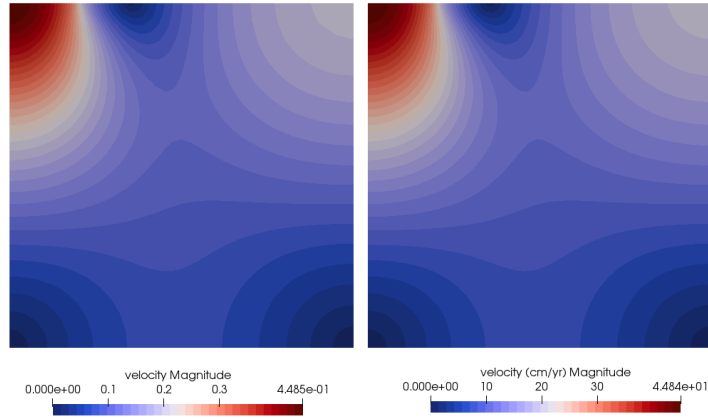
Left: ASPECT; Right: stone 64

It was run with ASPECT: prm file in stone 64 results folder.

25.3 Numerical Benchmarks

25.3.1 Bending of elastic slab (Gerya's book)

The sinking slab benchmark consists of a beam of elastic material which is placed in a weak and viscous surrounding medium. The initially unstressed beam is attached to the left domain boundary through boundary conditions. A stress is then applied to the beam in the form of gravity. The applied gravity force results in the deformation of the beam through bending. After 20 kyr, the gravity field is turned off and the elastic properties of the beam will then force itself to its original position. The set-up of the benchmark is given in the following figure:
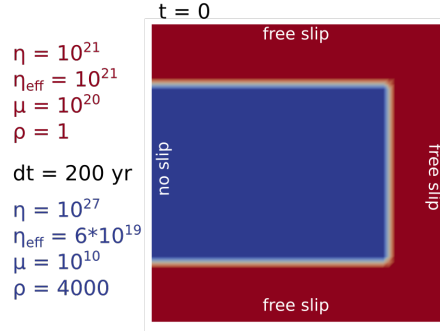
Set-up of the benchmark from [483]. The properties of the two materials are given on the left, together with the initial configuration of the benchmark.

The beam is surrounded by a low-density, low-viscosity and high shear modulus medium of which the specifications are given in the following table. The boundary conditions of the domain consist of a no slip condition at the left boundary where the slab is attached and free slip boundary conditions along all other sides. The results are calculated on a grid with a resolution of 50x50 elements containing 64 randomly distributed markers at startup. The time step is set to $\delta t = 200yr$ (i.e. gravity is switched off after 100 time steps).

Material properties	Elastic slab (fluid 1)	Surrounding medium (fluid 2)
Density ρ [kg/m^3]	4000	1
Viscosity η [Pa· s]	10^{27}	10^{21}
Shear modulus μ [Pa]	10^{10}	10^{20}
Maxwell time t_M [yr]	3.17Gyr	3.17×10^{-7}yr
eff. visc. η_{eff} [Pa·s]	6.307199602192306e+19	9.999999984145105e+20
visco-elasticity factor Z [-]	0.9999999369280039	1.5854895966744522e-09

25.3.2 Flexure of elastic plate (Choi et al)

This benchmark is presented in Choi, Tan, Lavier, and Calo [256] (2013). The setup is as follows:
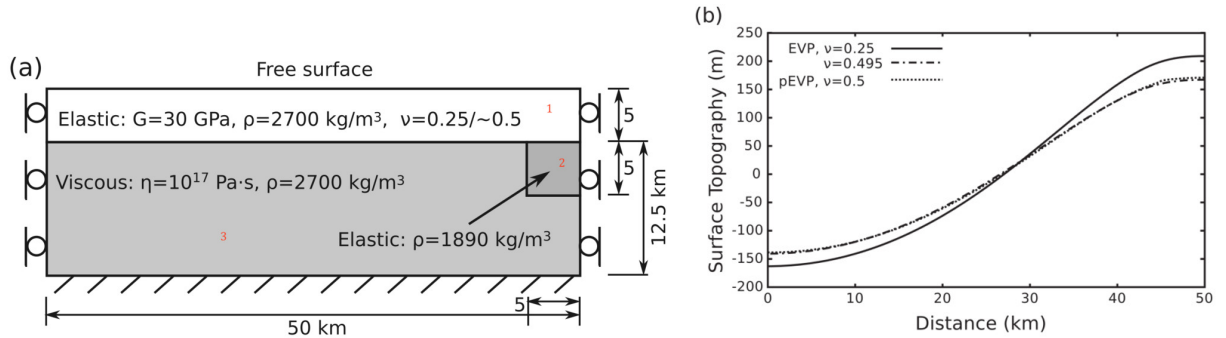
Taken from Choi, Tan, Lavier, and Calo [256]. "Effect of elastic compressibility on the prediction of an elastic thin plate subject to an uplifting load. (a) Model setup for a finite length elastic layer subjected to a finite length buoyant load applied on the bottom. (b) Profiles of mean-subtracted surface topography."

Material properties	elastic plate (1)	elastic block (2)	viscous mantle (3)
Density ρ kg m^{-3}	2700	1890	2700
Viscosity η Pa s	10^{35}	10^{35}	10^{17}
Shear modulus μ Pa	$30 \cdot 10^9$	$30 \cdot 10^9$	10^{50}
Maxwell time t_M yr	10569930	10569930	3.1709791983764584e-41
eff. visc. η_{eff} Pa s	4.73039776e+18	4.73039776e+18	1e+17

The value of $\eta_1 = \eta_2 = 10^{35}$ for the elastic materials was obtained through personal communication. The value of $\mu_3 = 10^{50}$ for the viscous material ensures that $\eta_{eff} = \eta_3$. Note that in

the publication the authors test both compressible and incompressible formulations, but we restrict ourselves to incompressible results since our code cannot handle compressible behavior. I also use dt=5year. Gravity is not specified in the paper.

The authors report a converged total relief of 306-308m.

This benchmark requires either a sticky air layer (see Section 16.1) on top of the plate or a deformable mesh (ALE formulation, see Section 16.1).

25.3.3 Elastic beam in viscous matrix - von Tscharner and Schmalholz

What follows is taken from von Tscharner and Schmalholz [1428] (2015).

This experiment first is a cylindrical elastic beam in a viscous matrix under gravity. The model box has the dimensions of $10x0.25x10$. The elastic beam which is vertically located in the middle of the model box has a thickness of $H = 1$ and a length of $L = 5$. The viscosity is $\eta_m = 1$ and $\eta_b = 10^{13}$ for the matrix and the beam, respectively. The elastic shear modulus is G m 5 10 10 and G b 5 10 3 for the matrix and beam, respectively, and the density difference is $\rho_b - \rho_m = 300$. These parameters provide a beam that is effectively elastic and a matrix that is effectively viscous. All model dimensions and material parameters are given in dimensionless numbers using the thickness of the beam H, the matrix viscosity and the background strain rate as characteristic parameters. The boundary conditions are free slip for all boundaries. The results of this simulation are shown in the figure below where the colors indicate the second invariant of the stress tensor. The elastic beam is deflected downward under vertical gravity (Figure b). When the gravity is turned off, the beam deflects upward due to the stored elastic energy and recovers the original rectangular shape which is stress free (Figure c). A similar test is given in Gerya's book [2010]. The test shows the reversible elastic deformation. The elastic beam recovers its original rectangular shape and stress state when the applied load is removed.
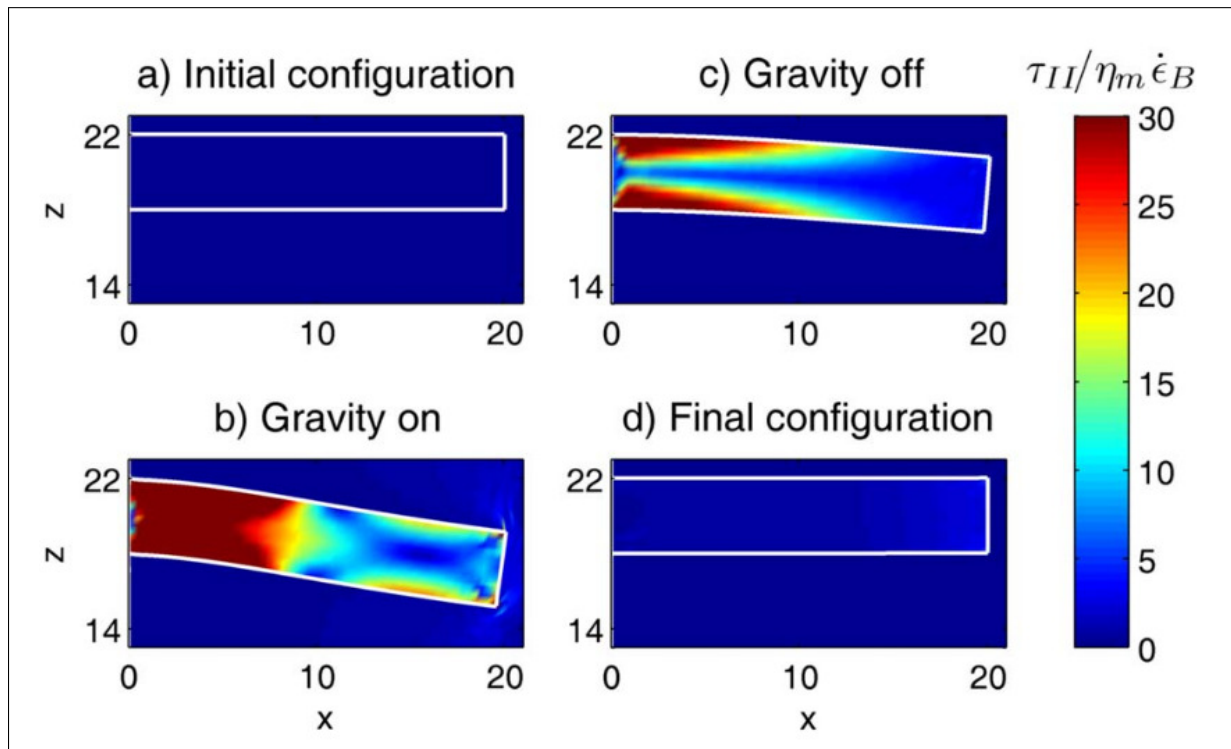
Reversible deformation of an elastic beam in a viscous matrix under gravity. (a) Unstressed initial configuration (gravity off). (b) Deformation of the elastic beam under gravity. (c) Gravity is turned off. (d) The elastic beam recovers the original rectangular shape with zero stress. Colors indicate the second invariant of the stress tensor τ_{II}.

25.3.4 Elastic beam in viscous matrix - Keller, May, and Kaus

This benchmark comes from Appendix B of Keller et al (2013) [728].

The domain is 7.5×5km. A dominantly elastic beam is fixed to, and protrudes horizontally from the left wall of the model box. Surrounding the elastic beam is a viscous, but inelastic fluid. All boundaries are free slip, except for the left wall, which is set to no slip in order to keep the bending beam fixed to the wall. The beam has a higher density than the surrounding fluid and thus will bend down elastically driven by gravity. After the beam has accumulated some elastic strain through bending down, gravity is switched off. If the stress evolution is implemented accurately, the elastic beam should now, free from the pull of gravity, move upwards again and restore its initial position.

Material properties are as follows:

- beam: $\rho = 1500$, $\eta = 10^{24}$, $\mu = 10^{10}$

- fluid: $\rho = 1000$, $\eta = 10^{18}$, $\mu = 10^{11}$

This choice of parameters leads to a Maxwell time $t_m = 0.32$ yr for the background fluid and Maxwell times of $t_m = 3.2$ Myr for the beam, meaning that the deformation in this benchmark problem, which occurs on a timescale of thousands to a million years, will lead to dominantly viscous deformation in the fluid, and dominantly elastic behaviour of the beam.

Keller et al set the numerical resolution to 300×200 elements, with 16 markers per elements for stress advection. Such a resolution is not feasible with our simple python implementation so the resolution is then set to 96×64.

The following plot comes from [728]:
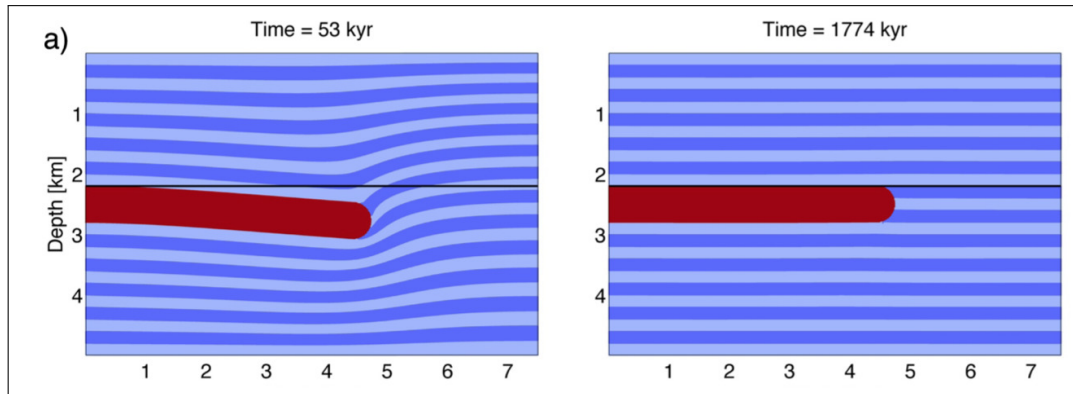
The elastic timestep is set to $\delta t_e = 100$yr and the tectonic timestep is set to the same value. This yields $\eta_{eff} = 10^{18}$ in the fluid, and $\eta_{eff} \simeq 3.15 \times 10^{19}$. After 50kyr, the gravity ($|\vec{g}| = 10$) is switched off and the model is ran for another 500kyr.

25.3.5 Boxcar load on an incompressible viscoelastic lithosphere - Wu (1992)

fully incompressible

This originates in Wu [1473] (1992). The domain is 2500 km long and it is either a halfspace in the vertical direction or a channel of 100 km width. The material is characterised by $\rho = 3400$ kg m^{-3}, Young's modulus $E = 1.13 \times 10^{11}$ Pa and a Poisson ratio $\nu = 0.5$ with $|\vec{g}| = 9.82$ m s^{-2}. The author explores the effect of a linear viscous mantle vs dislocation creep. Time evolution/relaxation figures are available.

Heaviside Response of ViscoElastic Channel to Boxcar Load
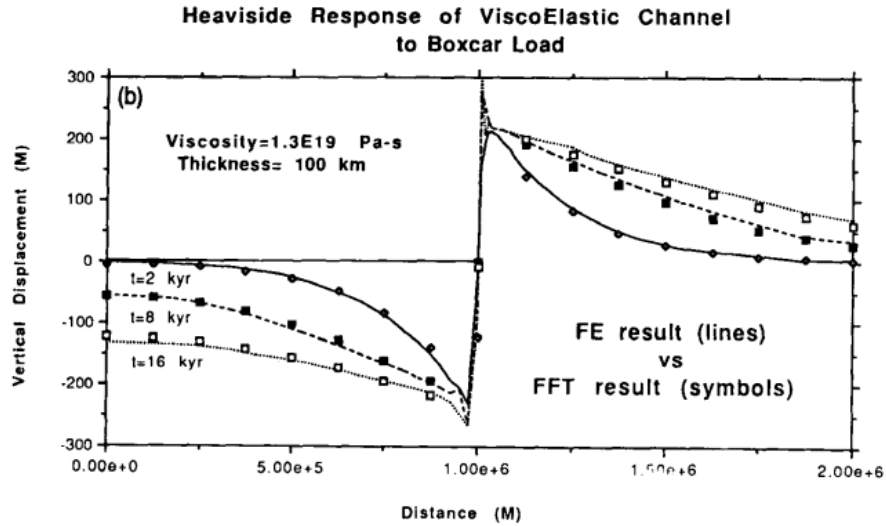
The Heaviside responses of linear viscoelastic earth models calculated with the finite element method (lines) are compared to those computed using the conventional transform method (symbols). For a 100 km thick channel.

also explore Wu & Peltier [1474] (1982).

25.3.6 Boxcar load on an incompressible viscoelastic lithosphere - Hampel et al. (2019)

Hampel, Lüke, Krause, and Hetzel [558] (2019)

fully incompressible
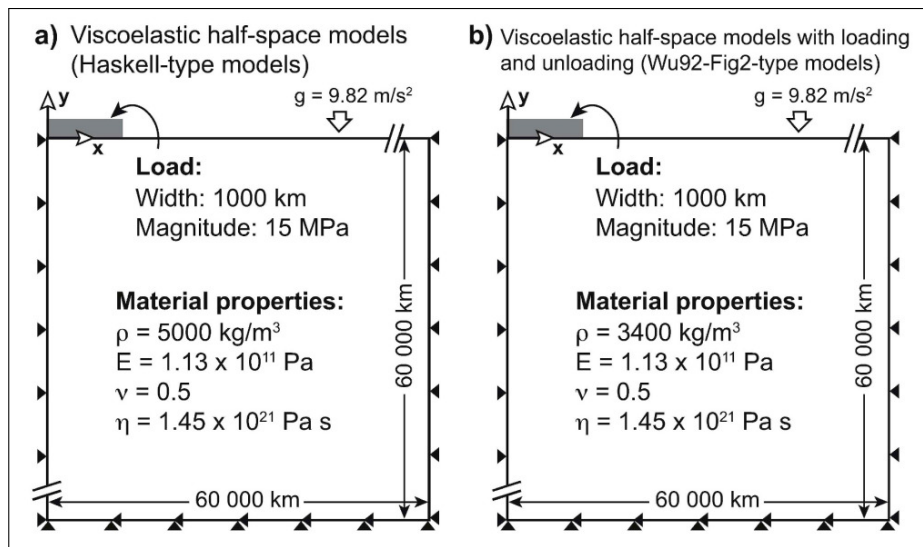
a) Viscoelastic Haskell-type half-space models. b) Viscoelastic half-space models used for modelling loading and subsequent unloading (after Wu, 1992; his Fig. 2). Both model types are meshed with 25x25 km large linear, rectangular plane strain elements suitable for incompressible materials. The same mesh and element type are used in all model runs. All viscoelastic half-space models have the same boundary conditions (indicated by black triangles): the model bottom is fixed in both the vertical and horizontal direction while the model sides are fixed in the horizontal direction. In models with an elastic foundation (Table 1), it is applied to the model surface. Abbreviations for model parameters are ρ density, E Young's modulus, ν Poisson's ratio, η viscosity and g acceleration due to gravity. Following Wu (1992), we use a value of $g = 9.82 m/s^2$ in the viscoelastic half-space models. In all models, the load is applied (and removed, if applicable) instantaneously. The magnitude of the applied load is 15 MPa, which is equivalent to about 1.5 km of ice. The left end of the load coincides with the origin of the coordinate system at the beginning of the model run. See text for details.

25.3.7 Kusznir and Bott (1977) experiment

The domain is a $2000\,\text{km} \times 80\,\text{km}$ Cartesian box. However since there is a vertical axis of symmetry it is then reduced to $1000\,\text{km} \times 80\,\text{km}$. It consists of two layers: the top layer is $20\,\text{km}$ thick and is purely elastic characterised by a Young's modulus $E = 1.7 \times 10^{11}\,\text{N}\,\text{m}^{-2}$ and a Poisson ratio $\nu = 0.25$. The lithosphere below is then $60\,\text{km}$ thick and is characterised by an elasto-viscous rheology. The elastic parameters are identical to the upper layer while the viscosity is set to $\eta_0 = 1 \times 10^{23}\,\text{Pa}\,\text{s}$.

We arbitrarily design the right side as being the axis of symmetry and the boundary condition on that side are then free slip. On the left side a uniform horizontal stress T is applied at time t=0. The viscous drag of the asthenosphere below is neglected and the authors do not mention the bottom of the domain deforming so we'll also impose free slip boundary conditions at the bottom.

Fig. 4 of the paper shows the vertical displacement of the upper layer so we will use a free surface boundary condition at the top. The model is ran for $1\,\text{Myr}$.
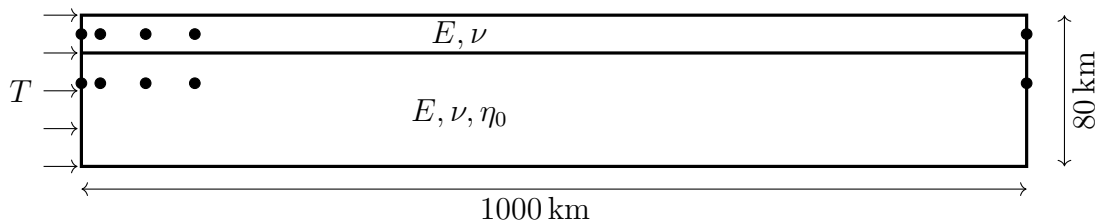
After a thorough read of the paper, I have noticed quite a few problems:

- the paper is old and has been digitized but the figures are missing a lot of lines/shades/points ... this could be remedied by finding the article in a library.

- in the intro it is stated: "Here we investigate the response of a lithosphere divided into upper elastic and lower uniform visco-elastic layers to simple boundary force and body force systems." The authors later talk about 'isostatic forces' opposing flexure. This means that buoyancy forces should be taken into account but there is no information about densities or gravity values!

- little uncertainty about boundary conditions. is it free slip or no-slip on the right side ? Looking at fig 4, it looks like the vertical displacement is zero at x=1000km?

- a Maxwell elasto-viscous rheology is used but this is only mentioned in the Appendix

- the dimensions of the thinned or thickened areas is simply not mentioned.

- Fig 1 shows triangular elements. No mention is made of resolution, type of element, ndofs, resolution tests, any numerical detail whatsoever. Given the age of the paper, I would guess P_1 elements.

- In the appendix they equate the viscous strain rate to $\sigma/4\eta$. Why 4 ?

- it is also not clear whether the domain is ALE or fully Lagrangian: does it shorten?

- the value of T is never specified!

- the paper was published 45 years ago, it is extremely unlikely any of the two authors is still available

Fig. 3. The horizontal lithospheric stresses for the model with uniform elastic layer thickness, as a function of time after the application of the stress T at the end of the lithosphere plate. Stresses in the elastic layer at 10 km depth (solid curve) and in the viscoelastic layer at 35 km depth (broken curve) are given at the following distances from the end of the lithosphere: (a) 0 km, (b) 20 km, (c) 60 km, (d) 125 km, (e) 1000 km (as shown on Fig. 2).

Fig. 4. Vertical displacement of the upper surface of the uniform layer thickness lithosphere model at time 0.5 My as a function of distance from the end of the lithosphere. Stress T applied at the end of the plate is equal to 0.1 kbar.

Taken from Kusznir and Bott [780]. Measurement locations are indicated on the setup figure above. probably should be reversed

25.3.8 Parallel-Plate Viscometer Problem - SNAC manual

A parallel-plate viscometer problem is simulated, in which viscoelastic material is squeezed between two parallel plates. The plates are moving at a constant velocity, v_0. Each plate has the length of $2L$ and is at a distance $2h$ from the other. No slip is assumed between the material and the plates. The approximate analytical solution is given in the book by Jaeger [665] (1969).

Model Setup: $L = 10$ m, $h = 5$ m, viscosity $\eta = 10^9$ Pa s, bulk modulus $K = 1.5$ GPa, shear modulus $\mu = 500$ MPa, $v_0 = 10^{-4}$ m s^{-1}, $dt = 1$ s (results compared after 500 time steps), mesh size: 20×10 m, each element is a 1 m cube.

Due to the assumption of the original problem setup, artificial forces should be added to left and right surfaces.
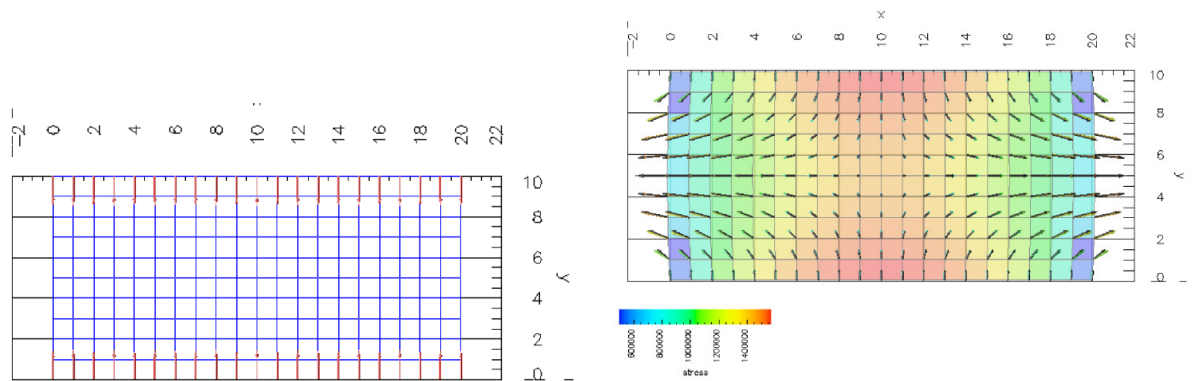
Left: The initial mesh (blue) with the velocity boundary condition (red arrows); Right: The second invariant of stress and velocities plotted on the deformed mesh. Colored arrows are for SNAC's solution, black ones for the analytic solution.

1184

$K = \lambda + \frac{2}{3}\mu$ so $\lambda = \frac{3500}{3} MPa$?

From wikipedia[1]

$$\nu = \frac{3K - 2\mu}{2(3K + \mu)} = \frac{4500 - 2*500}{2(4500 + 500)} = \frac{3500}{10000} = 0.35$$

so we find that the material is compressible $\nu = 0.35$

$$E = \frac{9K\mu}{3K + \mu} = \frac{9 * 1500 * 500 MPa^2}{4500 + 500 MPa} = \frac{9 * 1500 * 500}{5000} MPa = 1350 Mpa$$

25.3.9 Relaxation after extention - Hassani syllabus

compressible $\nu = 0.25$

Un essai de relaxation consiste à imposer à un instant donné une déformation que l'on maintient constante par la suite. On observe alors comment évolue la contrainte au cours du temps.

The setup of the experiment is shown in the following figure:
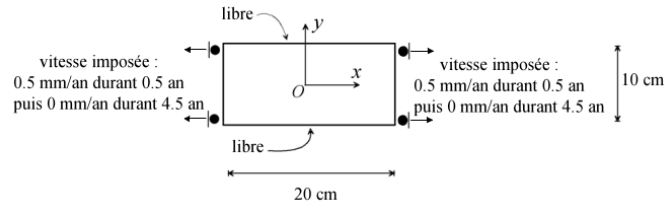
The total duration of the experiment is $T = 5\,\text{yr} \simeq 15.75 \times 10^7\,\text{s}$. The duration of the loading is $T/10 = 6\,\text{months}$ while the duration of the subsequent relaxation is then $9T/10 \simeq 4.5$ yr.

The loading velocity is $v = 1\,\text{mm}\,\text{yr}^{-1} \simeq 3.17 \times 10^{-11}\,\text{m}\,\text{s}^{-1}$. The sample has size $L \times L/2 = 20 \times 10$cm and the strain rate is then $v/L \simeq 1.5 \times 10^{-10}\,\text{s}^{-1}$. Young's modulus is set to $1 \times 10^{11}\,\text{Pa}$ and the Poisson ratio is 0.25, i.e. $\mu = 40$ GPa. The viscosity is set to $\eta_0 = 1 \times 10^{18}\,\text{Pa}\,\text{s}$. The Maxwell time is then $t_M = \eta/\mu = 0.8$ yr, which is also the time it takes to reduce the maximum stress by a factor e.
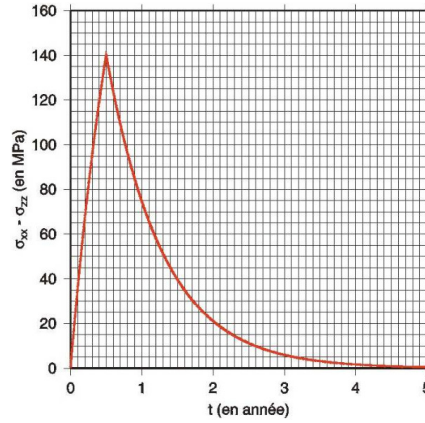
[1]https://en.wikipedia.org/wiki/Lame_parameters

25.3.10 Role of elasticity in slab bending - Fourel, Goes, and Morra [432] (2014)
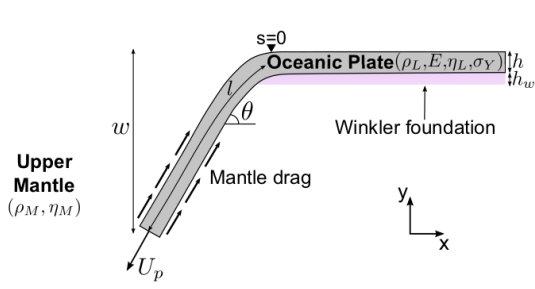
Figure 1. Schematic of the model setup and definition of parameters. The slab of thickness h, density ρ_L, Young's modulus E, viscosity η_L, and in some cases yield stress σ_Y is freely sinking into an infinite mantle of density ρ_M and viscosity η_M, achieving a subduction velocity U_p, deflection w and slab dip θ. Slab deformation is modeled in 2-D, but the analytically calculated mantle drag that is applied along the slab's sides and tip is for 3-D flow around a finite-width plate. The plate is supported in the top few kilometers by a Winkler foundation of thickness h_w. The top surface and the trailing edge are free.

25.3.11 Shear test in 2D - Farrington, Moresi, and Capitanio [412] (2014)

This experiment consists of a viscoelastic material undergoing simple shear at a constant rate from a time t_0 though to t_{max}. At t_{max}, the shearing velocity is taken to zero using a no-slip velocity boundary condition. The viscoelastic st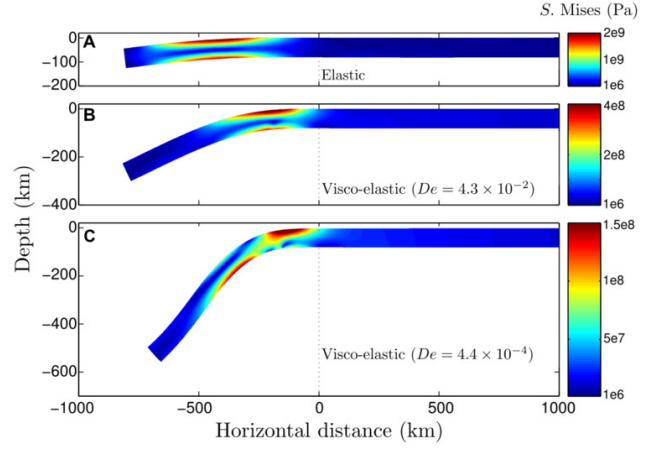resses then decay with time while the material deformation rate remains zero. The xy component of stored stress, i.e. the nonviscous portion of the total stress, is given by

$$
\begin{aligned}
\tau_{xy}^{stor} &= \exp-\frac{\mu}{\eta}t\left(C_2\cos\left(\frac{Vt}{h}\right)-C_1\sin\left(\frac{Vt}{h}\right)\right)-C_2 && \text{if } t < t_{max} \\
&= \left[\exp-\frac{\mu}{\eta}t_{max}\left(C_2\cos\left(\frac{Vt}{h}\right)-C_1\sin\left(\frac{Vt}{h}\right)\right)-C_2\right]\exp-\frac{\mu}{\eta}(t-t_{max}) && \text{if } t > t_{max} \quad (25.6)
\end{aligned}
$$

where V is the shear velocity along the top wall boundary, h is the height of the box,

$$
C_1 = -\frac{V^2\eta^2\mu}{\mu^2h^2+V^2\eta^2}
$$

$$
C_2 = -\frac{Vh\eta\mu^2}{\mu^2h^2+V^2\eta^2}
$$
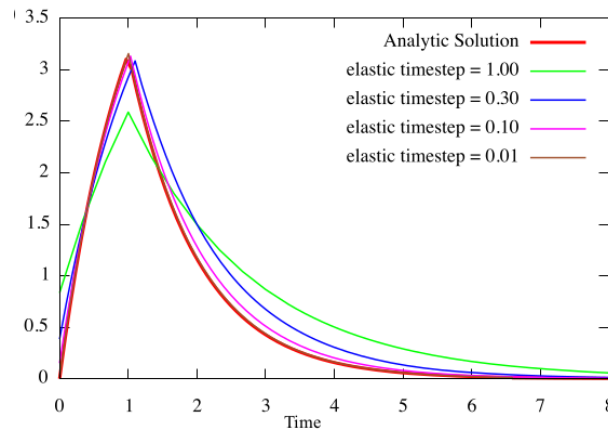
The xy component of dimensionless stored stress with dimensionless time for the 2-D viscoelastic material undergoing simple shear and relaxation. Nondimensional material parameters of $\eta = 10^2$, $\mu = 10^2$, $t_M = 1$, $V = 0.05$, $h = 1$ with $\Delta t_e \in [0.01, 1]$ and $\Delta t_c = \frac{1}{3}\Delta t_e$.

Figure above shows the resulting xy component of the deviatoric stored stress term for a 2-D viscoelastic material across a range of Δt_e. At longer Δt_e, the stored stress appears under resolved, indicating these larger elastic time step values capture dynamics that occur on time scales approximately equal to or greater than the Maxwell relaxation time, that is with a portion of viscous deformation. For shorter Δt_e, the numerically calculated stress approaches the analytic solution, indicating that the elastic time step is sufficiently small to fully capture the elastic stored stresses produced within the material under the applied strain rate.

25.3.12 Tortion test in 3D - Farrington, Moresi, and Capitanio [412] (2014)

insert here eq 9 of paper
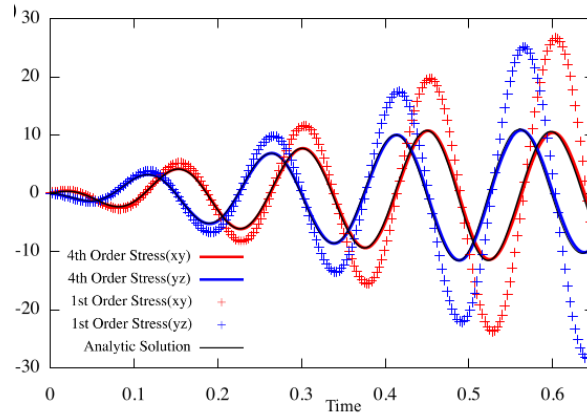
The analytic solution outlined by equation (9) can be extended from this essentially 1-D test into 3-D by applying the shear velocity in the x-z plane. Testing of the full viscoelastic implementation including the rotation terms is possible by placing this 3-D shear test in a coordinate system under going solid body rotation.

The xy and yz stress components of a material undergoing simple shear within a 3D rotating reference frame. Nondimensional material parameters of $\eta = 100$, $\mu = 100$, $\alpha = 1$, $V = 0.3$, $h = 1$, $t_{max} = 0.5$ and $\omega = 42$. Note the coordinate system for this test has the y axis in the vertical with the z axis in plane. Results for a first (crosses) and fourth- (lines) order accurate Jaumann stress rate integration scheme are shown in comparison to the analytical solution (black) given by equation (9) within the rotating frame.

Figure above shows the evolution of the stress in comparison to the analytical solution of equation (9) placed within the rotating frame. The rotating frame is achieved by imposing a velocity boundary condition of a constant solid body rotation about the y axis in addition to the shearing rate. The stress within this rotating frame is given by equation (9), with the stress in the nonrotating frame found by applying a rotation matrix, R, to the nonrotating stress solution. That is, $\tau = R^{-1}\tau' R$, where $'$ denotes the rotated frame, R is the rotation matrix about the y axis given by

$$R = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}$$

with $\theta = \omega t$, ω being the dimensionless rotation rate. The stress components shown in Figure 1b from within the nonrotating frame can then be given by $\tau_{xy}^{stor} = \tau_{xy}^{stor'} \cos\theta$ and $\tau_{yz}^{stor} = \tau_{yz}^{stor'} \sin\theta$. It can be seen that, using a higher-order Jaumann stress rate advection scheme results in accurate stress advection and rotation within the full 3-D space plus time domain. It should be noted here that the velocity field used in this test was chosen to rigorously test the rotational terms in equation (6). Whether these rotational terms are required for individual models is dependent upon the model setup. For subducting slabs at a constant curvature, the individual parcels of material experience purely rotational effects, accounting for this within the stress history term would then be required for consistency.

finish!

1187

## 25.3.13 Cylindrical tunnel - Segall book (?)

compressible $\nu = 0.25$

communicated to me by L .van de Wiel.

This is a cylindrical tunnel (2D magma chamber) in an infinite space, with certain radius. The material close around the hole is warm, and and such viscous. The material further from the hole is cold, and purely elastic. (accomplished by setting and absurdly high viscosity) There is a clear transition radius between the two properties The hole contains a pressure, causing the space to expand.

After the initial elastic deformation, viscous deformation continues in the viscous region of the domain. I have the implementation of the analytical solution attached to save you time.

See plot with radial deformation for t=0 (green), t=300s (red), t=600s (blue), and t=3000s (black). thin line is analytic, points are numeric. (with parameters and size parameters as in analytic.f)
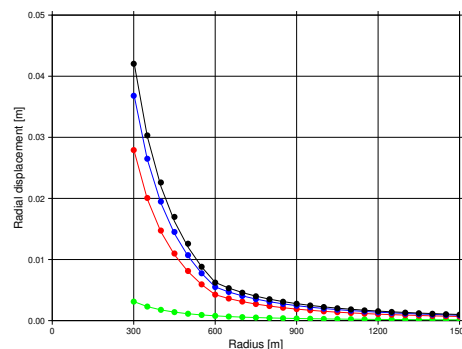
25.3.14 Relevant literature & various notes

- convection of viscoelastic fluids: Harder [579], Moser, Matyska, Yuen, Malevsky, and Harder [962], Zhong, Gurnis, and Moresi [1519], Moresi, Dufour, and Mühlhaus [952], Mühlhaus and Regenauer-Lieb [969], Li and Khayat [830], Li and Khayat [831], Furuichi, Kameyama, and Kageyama [455]

- stress buildup associated with viscoelastic rheology Kusznir and Bott [780], Kusznir and Park [781], Poliakov, Cundall, Podlachikov, and Lyakhovsky [1072], Marques and Podladchikov [884]

- viscoelastic effects on geodynamical pbs involving gravitational instability Poliakov, Cundall, Podlachikov, and Lyakhovsky [1072], Kaus and Becker [721], Burov and Molnar [199], and Schmeling et al. [1200], Hanyk, Moser, Yuen, and Matyska [578]

- large strain eulerian viscoelasticity Schmalholz, Podladchikov, and Schmid [1193], Vasilyev, Podladchikov, and Yuen [1412], Cooper, Lenardic, Levander, Moresi, and Benn [297], Moresi, Quenette, Lemiale, Mériaux, Appelbe, and Mühlhaus [954], Furuichi, Kameyama, and Kageyama [455], and Popov and Sobolev [1076]

Farrington, Moresi, and Capitanio [412] states "Funiciello et al. [2003] implemented a viscoelastic rheology in numerical models of subduction, performing a range of numerical simulations to investigate its effect on subducting slab dynamics. Similar methodologies have addressed the details of viscoelastic stress within the bending zone during subduction, although a comparison between viscous and viscoelastic rheology was lacking [Capitanio et al., 2009; Capitanio and Morra, 2012]. Muhlhaus and Regenauer-Lieb [2005] and Moresi et al. [2002] have studied the role of elasticity in

mantle convection, comparing the viscous case to that of viscoelastic, and Kaus and Becker [2007] discussed the effect of elasticity on layered Rayleigh-Taylor instabilities. However, a systematic study into the effects of elastic stresses on models of free subduction has yet to be completed. Morra and Regenauer-Lieb [2006], Funiciello et al. [2003], Capitanio et al. [2007], Yamato et al. [2007], and Royden and Husson [2006] have included a viscoelastic slab in subduction models, without explicitly studying the effects of the elastic component across a range of parameters."

Check early paper by Braun & Beaumont (1987) [150]

Asgari and Moresi [32] (2012) Herwegh, Poulet, Karrech, and Regenauer-Lieb [599] (2014) Dansereau, Weiss, Saramito, and Lattes [324] (2016) Thielmann, Kaus, and Popov [1339] (2015) Beuchert, Podladchikov, Simon, and Rüpke [93] (2010) Sanan, May, Bollhöfer, and Schenk [1182] (2020) von Tscharner and Schmalholz [1428] (2015) Naliboff, Lithgow-Bertelloni, Ruff, and Koker [983] (2012) Peltier [1048] (1974)

Chapter 26

Geophysical data

26.1 Earth: The PREM model

The density profile Let us define $x = r/R$. Following table I of Dziewonski & Anderson (1981) [380] we have

- for the inner core $0 < r < 1221.5$km (or $0 < x < 0.19172814314$):

$$\rho(x) = 13.0885 - 8.8381x^2$$

- for the outer core $1221.5 < r < 3480$km (or $0.19172814314 < x < 0.5462250824$:

$$\rho(x) = 12.5815 - 1.2638x - 3.6426x^2 - 5.5281x^3$$

- for the Lower mantle $3480 < r < 5701$km:

$$\rho(x) = 7.9565 - 6.4761x + 5.5283x^2 - 3.0807x^3$$

- for the transition zone 1 $5701 < r < 5771$km:

$$\rho(x) = 5.3197 - 1.4836x$$

- for the transition zone 2 $5771 < r < 5971$km:

$$\rho(x) = 11.2494 - 8.0298x$$

- for the transition zone 3 $5971 < r < 6151$km:

$$\rho(x) = 7.1089 - 3.8045x$$

- Low velocity zone $6151 < r < 6291$km:

$$\rho(x) = 2.6910 + 0.6924x$$

- LID $6291 < r < 6346.6$km:

$$\rho(x) = 2.6910 + 0.6924x$$

- Lower Crust $6346.6 < r < 6356$km:

$$\rho(x) = 2.9$$

- Upper Crust $6356 < r < 6368$km:
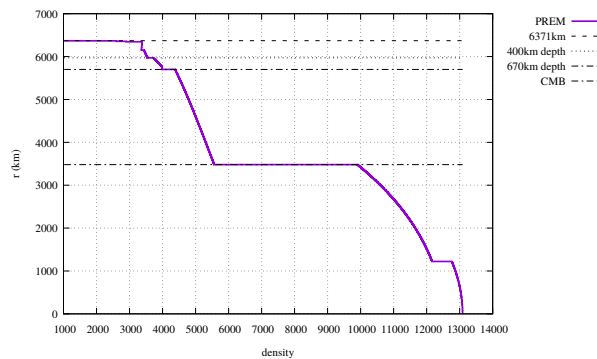
$$\rho(x) = 2.6$$

- Ocean $6368 < r < 6371$km

$$\rho(x) = 1.020$$

Note that the returned densities should be multiplied by 1000 to obtain units of kg/m^3.
    One can verify that the functions above yield the familiar PREM density profile:

1191

Corresponding gravity field Following Eq. (17.39), the radial component of the gravitational acceleration at a position r outside of the Earth is given by:

$$g_r(r) = -\frac{1}{r^2} \int_0^r 4\pi\mathcal{G}\rho(r')r'^2 dr' = -\frac{4\pi\mathcal{G}}{r^2} \int_0^r \rho(r')r'^2 dr' \qquad (26.1)$$

This integral can be broken up into layer integrals and we can compute the contribution of each layer to the gravity value $g_r(r)$. A layer is characterised by r_{in} and r_{out}. Inside r_{in} the gravity is zero. Between r_{in} and r_{out} the integral runs from r_{in} to $r \leq r_{out}$. Anout outside the layer the integral runs from r_{in} to r_{out}.

For simplicity the following integrals are computed for $x \in [0,1]$:

- inner core

$$
\begin{aligned}
inside \quad g_{ic}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int_0^x (13.0885 - 8.8381x'^2)x'^2 dx' \\
&= \frac{4\pi\mathcal{G}}{x^2}(4.36283x^3 - 1.76762x^5) \\
&= 4\pi\mathcal{G}(4.36283x - 1.76762x^3) \qquad (26.2) \\
outside \quad g_{ic}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int_0^{0.19172814314} (13.0885 - 8.8381x'^2)x'^2 dx' \\
&\simeq 0.0302907\frac{4\pi\mathcal{G}}{x^2} \qquad (26.3)
\end{aligned}
$$

- outer core

$$
\begin{aligned}
inside \quad g_{oc}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int_{0.19172814314}^x (12.5815 - 1.2638x' - 3.6426x'^2 - 5.5281x'^3)x'^2 dx' \\
&= 4\pi\mathcal{G}(-0.92135x^4 - 0.72852x^3 - 0.31595x^2 + 4.19383x - 0.028896/x^2) \qquad (26.4) \\
outside \quad g_{oc}(x) &= \frac{4\pi\mathcal{G}R^3}{x^2} \int_{0.19172814314}^{0.5462250824} (12.5815 - 1.2638x' - 3.6426x'^2 - 5.5281x'^3)x'^2 dx' \\
&\simeq 0.56663\frac{4\pi\mathcal{G}}{x^2} \qquad (26.5)
\end{aligned}
$$

- lower mantle

$$
\begin{aligned}
inside \quad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int^x ()x'^2 dx' \\
outside \quad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int ()x'^2 dx'
\end{aligned}
$$

$$(26.6)$$

- transition zone 1

$$
\begin{aligned}
inside \quad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int^x ()x'^2 dx' \\
outside \quad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2} \int ()x'^2 dx'
\end{aligned}
$$

$$(26.7)$$

- transition zone 2

$$
\begin{aligned}
inside \qquad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2}\int^x ()x'^2dx' \\
outside \qquad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2}\int ()x'^2dx'
\end{aligned}
$$

(26.8)

- transition zone 3

$$
\begin{aligned}
inside \qquad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2}\int^x ()x'^2dx' \\
outside \qquad g_{lm}(x) &= \frac{4\pi\mathcal{G}}{x^2}\int ()x'^2dx'
\end{aligned}
$$

(26.9)

$$
\begin{aligned}
g_{lm}(r) &= \int_{3480/6371}^{5701/6371} (7.9565 - 6.4761x + 5.5283x^2 - 3.0807x^3)x^2dx \simeq 0.904793\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{tz1}(r) &= \int_{5701/6371}^{5771/6371} (5.3197 - 1.4836x)x^2dx \simeq 0.0354823\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{tz2}(r) &= \int_{5771/6371}^{5971/6371} (11.2494 - 8.0298x)x^2dx \simeq 0.1026\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{tz3}(r) &= \int_{5971/6371}^{6151/6371} (7.1089 - 3.8045x)x^2dx \simeq 0.0892215\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{lvz} &= \int_{6151/6371}^{6291/6371} (2.6910 + 0.6924x)x^2dx \simeq 0.0705516\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{lid} &= \int_{6291/6371}^{6346.6/6371} (2.6910 + 0.6924x)x^2dx \simeq 0.0289968\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{lc} &= \int_{6346.6/6371}^{6356/6371} 2.9x^2dx \simeq 0.00425234\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_{uc} &= \int_{6356/6371}^{6368/6371} 2.6x^2dx \simeq 0.00488337\frac{4\pi\mathcal{G}R^3}{r^2} \\
g_o &= \int_{6368/6371}^{1} 1.020x^2dx \simeq 0.000480075\frac{4\pi\mathcal{G}R^3}{r^2}
\end{aligned}
$$

(26.10)

Finally

$$
\begin{aligned}
|g_r(r)| &= g_{ic}(r) + g_{oc}(r) + g_{lm}(r) + g_{tz1}(r) + g_{tz2}(r) + g_{tz3}(r) + g_{lvz}(r) + g_{lid}(r) + g_{lc}(r) + g_{uc}(r) + g_o(r) \\
&= \frac{4\pi\mathcal{G}R^3}{r^2}(0.0302907 + 0.56663 + 0.904793 + 0.0354823 + 0.1026 + 0.0892215 + \\
&\quad 0.0705516 + 0.0289968 + 0.00425234 + 0.00488337 + 0.000480075) \\
&\simeq \frac{4\pi\mathcal{G}R^3}{r^2}1.838181685
\end{aligned}
$$

At the surface of the Earth, $r = R$ so we arrive at (after multiplying by 1000, see comment above):

$$
\boxed{g_{PREM}(R) \simeq 4\pi \cdot 6.67408 \times 10^{-11} \cdot 6371 \times 10^3 \cdot 1.838181685 \simeq 9.82194}
$$

1193

All these calculations should be rechecked, although obviously the obtained value makes much sense.

26.2 Earth: From 1D tomography to density/temperature

The mantle is heterogeneous but it is also inaccessible. This means that one must rely on indirect methods to probe its structure. Seismic tomography is a technique for imaging the subsurface of the Earth with seismic waves produced by earthquakes or explosions. P-, S-, and surface waves can be used for tomographic models of different resolutions.

Seismic velocity is a meaningful parameter for the interior dynamics of the Earth because there exists a direct relation between seismic velocity and density. Such a relation was analysed experimentally by (for instance) Barton (1986) who used laboratory measurements of P-wave seismic velocity and density of rocks [55].

Fourty years later or so, a crucial question remains: what is the exact form of the relation between density and seismic velocity for the entire Earth's mantle?

I will here not go into the details of the underlying theories and their approximations but will show a few useful results.

From tomography to density, the workflow is usually as follows:

$$d \ln V_p \to d \ln V_s \to d \ln \rho \to d\rho$$

Note that if the method is based on shear wave tomography the conversion $d \ln V_p \to d \ln V_s$ is not necessary. Also the last step $d \ln \rho \to d\rho$ requires a background density field, often taken to be either the PREM model or AK135 (see Section 26.1).

On the following plots are shown radial averages of the ratio $d \ln V_s / d \ln V_p$ and/or $d \ln \rho / d \ln V_s$:
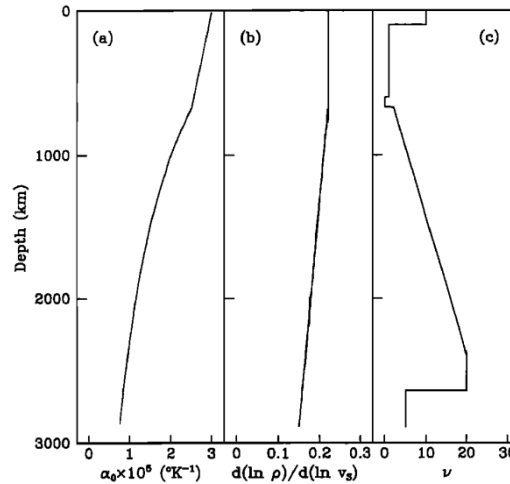
Taken from Pari and Peltier [1033] (1995).
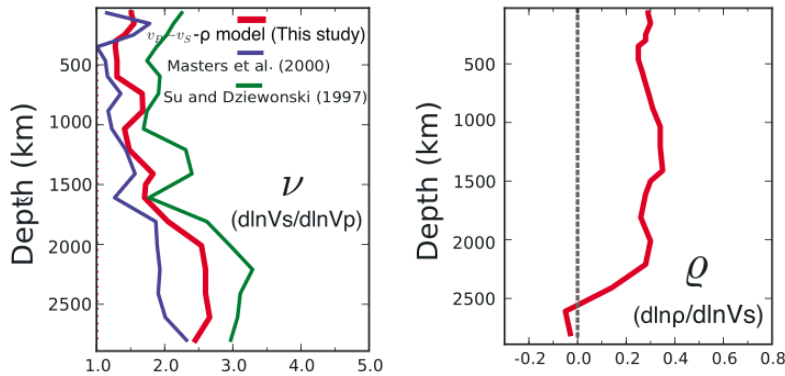
Taken from Moulik & Ekstrom (2016) [964]
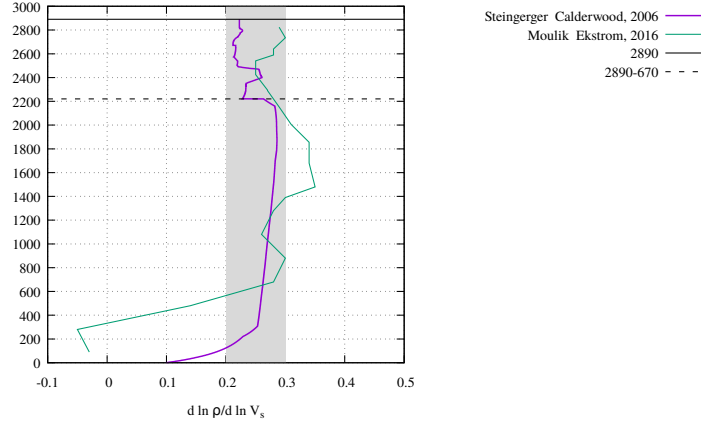
Profiles of scaling factor $\xi = d\ln\rho/d\ln V_s$. Data from Steinberger & Calderwood (2006) [1284] and Moulik & Ekstrom (2016) [964]. Data available in images/dlnrhodlnvs/

Then

$$\delta\ln(\rho(r,\theta,\phi)) = \xi(r)\cdot\delta\ln(V_s(r,\theta,\phi))$$

with

$$\delta\ln(\rho(r,\theta,\phi)) = \frac{\delta\rho(r,\theta,\phi)}{\rho_{ref}(r)}$$

where ρ_{ref} is a radial profile, PREM for instance in the case of S40RTS so finally

$$\delta\rho(r,\theta,\phi) = \rho_{ref}(r)\cdot\delta\ln(\rho(r,\theta,\phi)) = \rho_{ref}(r)\cdot\xi(r)\cdot\delta\ln(V_s(r,\theta,\phi))$$

Relevant Literature: [1155]

1196

26.3   Earth: radial viscosity profile
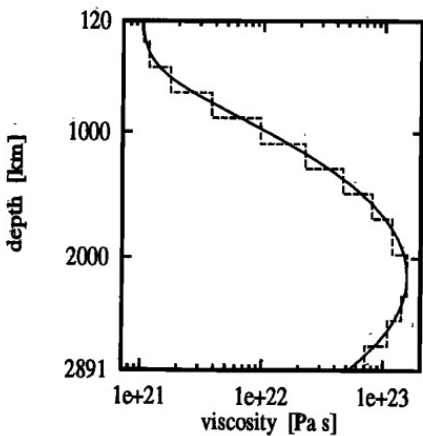
1995: Taken from Hanyk *et al.* (1995) [578]. $\eta(z) = (1 + 214.3z \exp -16.7(0.7 - z)^2) \times 10^{21}$pascals
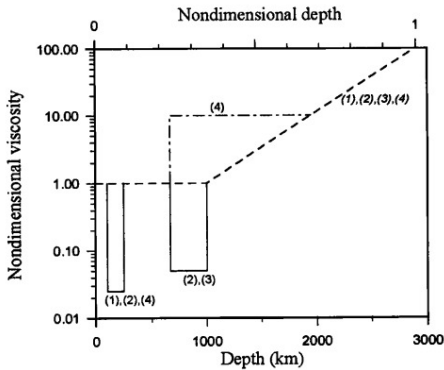
1997: Taken from Cserepes & Yuen (1997) [311]
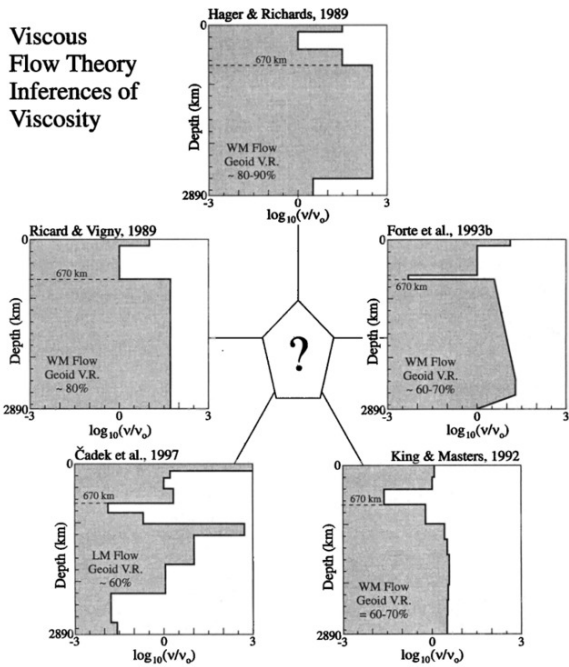
1998: Taken from Pari and Peltier [1032] (1998).
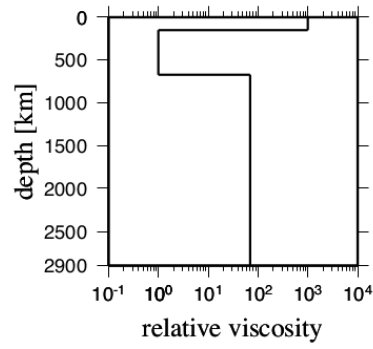
2001:Radial viscosity profile of the reference model. 3-layered model is adopted: the lithosphere (0 km to 150 km), the upper mantle (150 km to 670km) and the lower mantle (670 km to 2900 km). Taken from [1489]
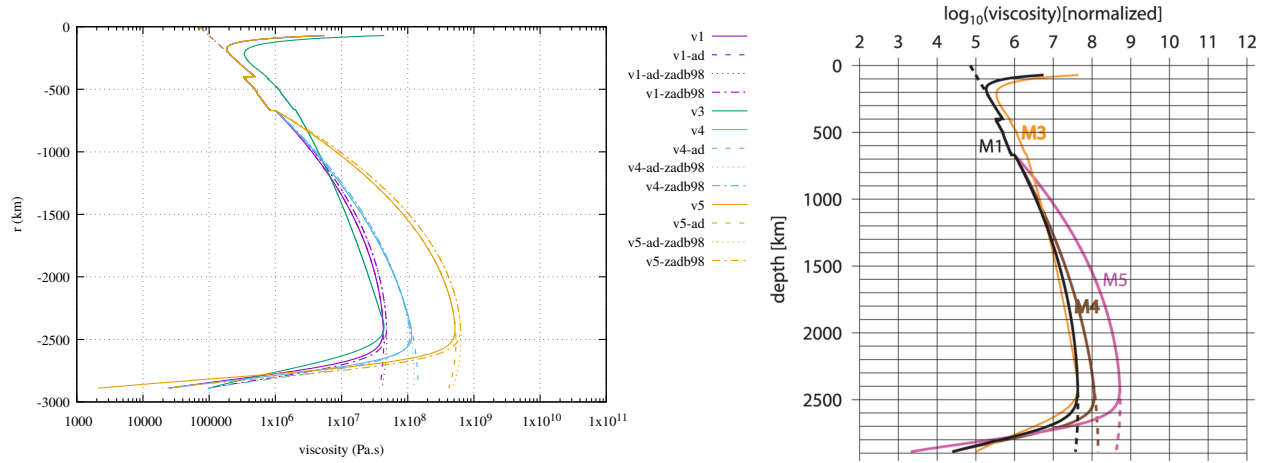
2006: Non-optimized, normalized viscosity profiles, as sent by B. Steinberger, corresponding to fig. 4 in Steinberger & Calderwood (2006) [1284].
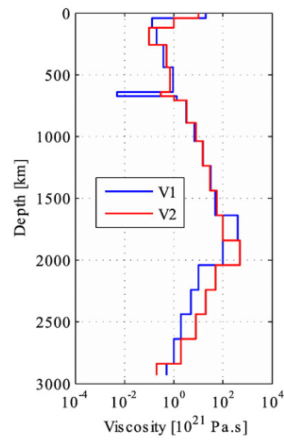
Taken from Moucha et al. [963] (2008).
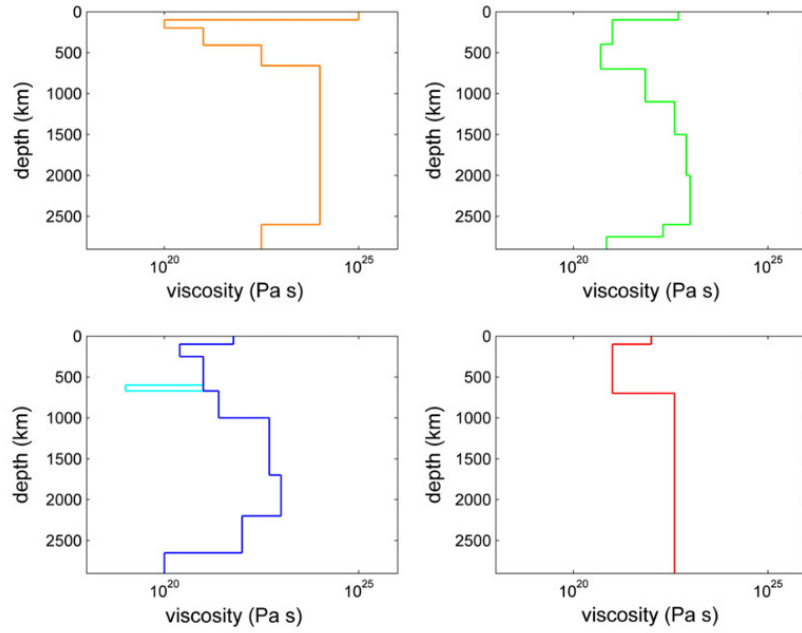
2011: Taken from Cadio *et al.* [215]. Mantle viscosity structure employed in calculating synthetic geoid anomalies. Red: VR (Ricard *et al.*, 1993); Blue: VMF (Mitrovica and Forte, 2004); Cyan: VMF-LVZ (Mitrovica and Forte, 2004); Green: VSC (Steinberger and Calderwood, 2006); Orange: VYN (Yoshida and Nakakuki, 2009).
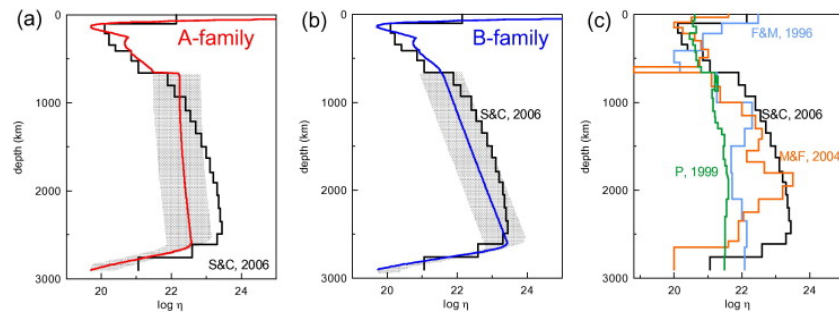
2012: Taken from Ciskova *et al.* [277].
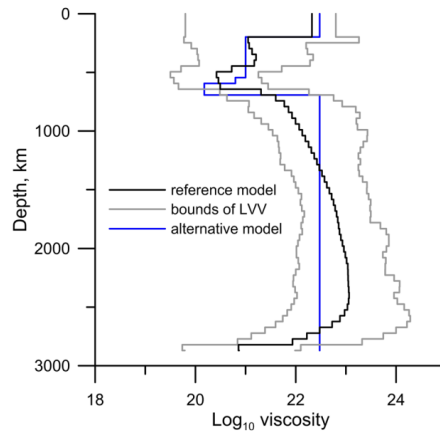
2014 Taken from Kaban *et al.* [698]. (Black) reference radial viscosity from (Steinberger and Calderwood 2006); (Blue) alternative viscosity model of Kaban and Trubitsyn (2012); (Light Grey) limits of viscosity variations in the model with LVV (Petrunin *et al.* 2013).

1199
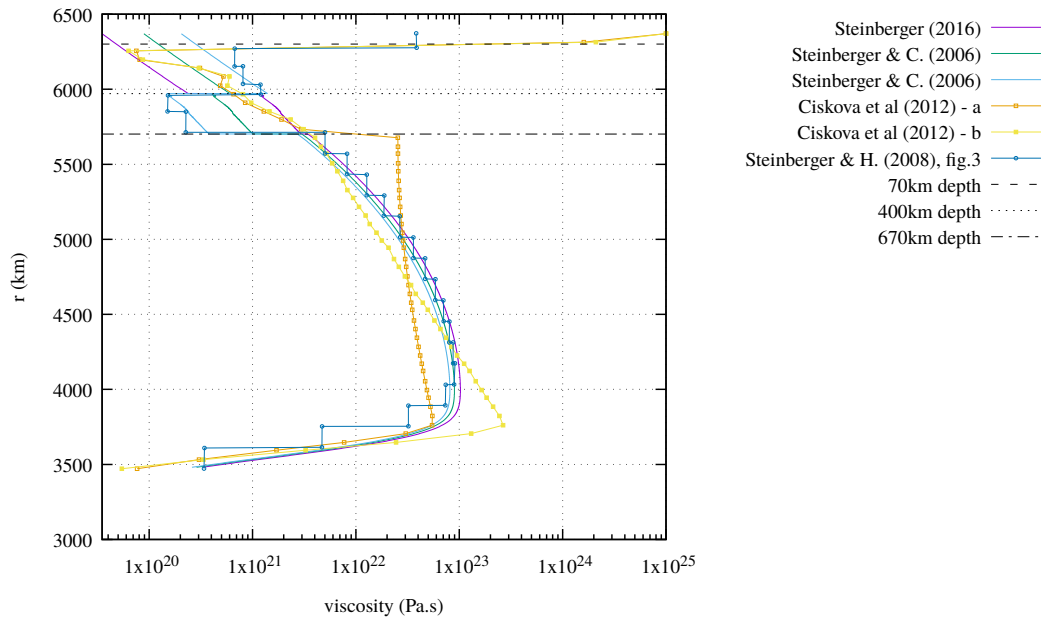
Steinberger (2016) [1285]; Ciskova *et al.* (2012) [277].
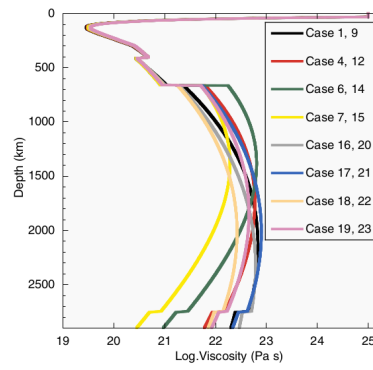
2019: Taken from Kaneko *et al.* [707] (2019).
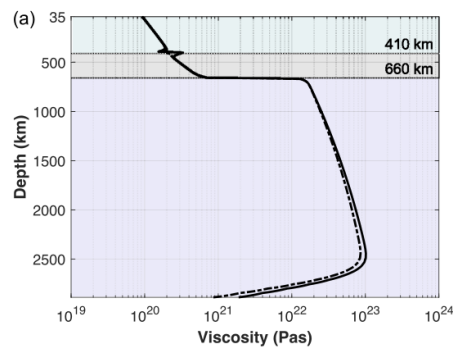
2023: Taken from Neuharth and Mittelstaedt [991] (2023).
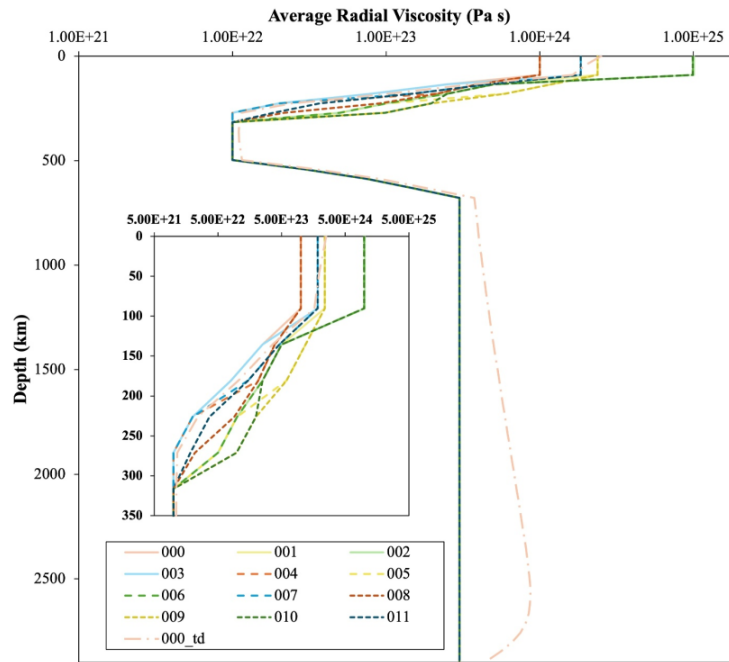
2024: Taken from Plimmer, Davies, and Panton [1068] (2024).

📥 Relevant Literature:

- Viscosity profile of the lower mantle [393]

- Matyska *et al.* (2011) [893]

- Flament (2019) [422]

- Mitrovica & Forte [929]

- King & Masters [748]

- Rudolph *et al.* [1162]

- steinberger & holme [1283]

- steinberger [1285]

- Cadek & Yuen [213]

- supp of [53]

26.4 Earth: radial temperature profile

The method of manufactured solutions is a relatively simple way of carrying out

We start by the first sentence of Bunge (2005) [184]: "The average temperature increase through Earth's crust and mantle is called the geotherm. Its basic form is assumed to consist of adiabatic regions where temperatures rise only slightly with depth, and of narrow thermal boundary layers where temperatures increase rapidly over a depth of a few hundred kilometers (Jeanloz and Morris, 1986) [680]."

Before we look further at the equations behind this temperature profile, we must look at the basic assumption we make about the type of convection taking place in the mantle: layered convection vs. whole mantle convection, as depicted here:

Taken from https://geologyengineering.com/2020/05/mantle-convection/
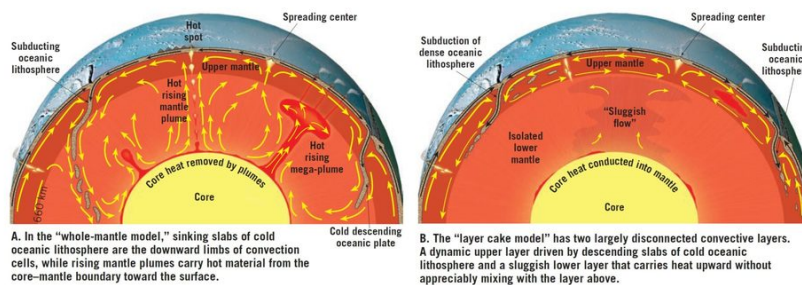
Indeed, the type of convection is then expected to have an strong influence on the radial temperature profile:
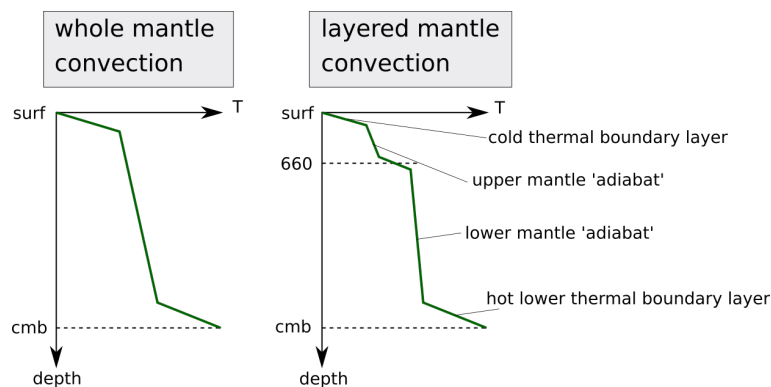
This is also to be found in Poirier's book [1071]:
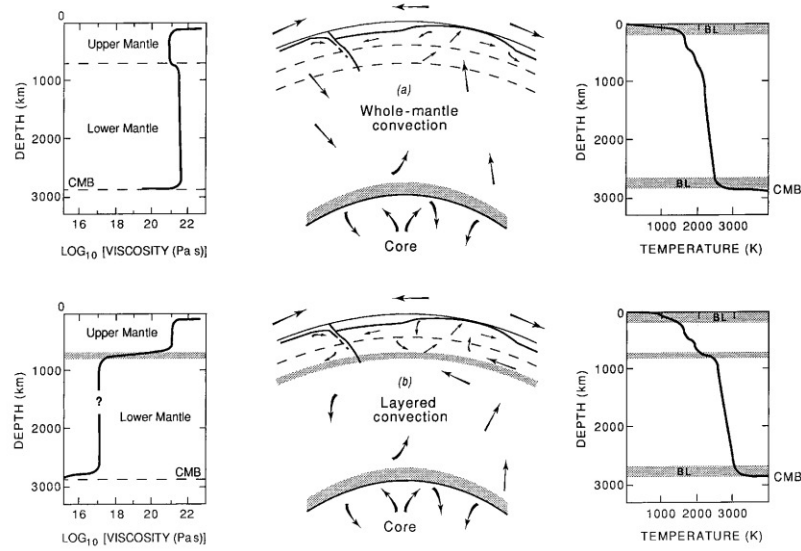
Schematic diagrams of (a) whole-mantle and (b) layered convection models, with corresponding temperature and viscosity profiles (after Peltier & Jarvis

(1982) [1050].)

The two-layered convection hypothesis relies essentially on the viscosity jump at the 660 discontinuity and seems to be generally accepted. It also forms the basis of Čadek & van den Berg (1998) [214] in which the authors carry out an inversion to obtain radial profiles of temperature and viscosity in the Earth's mantle inferred from the geoid and lateral seismic structure:
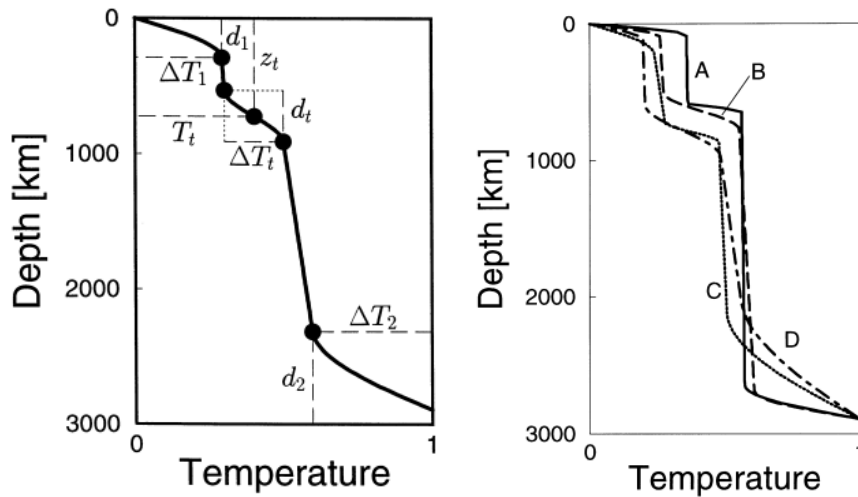
Taken from [214]. Left: Parameterization of the geotherm used in the paper; Right: Four model geotherms reducing the misfit by 70%. The differences

between the models illustrate uncertainties of the solution.

More recently, Katsura *et al.* (2010) [717] have constructed the following mantle temperature and temperature gradient profiles:
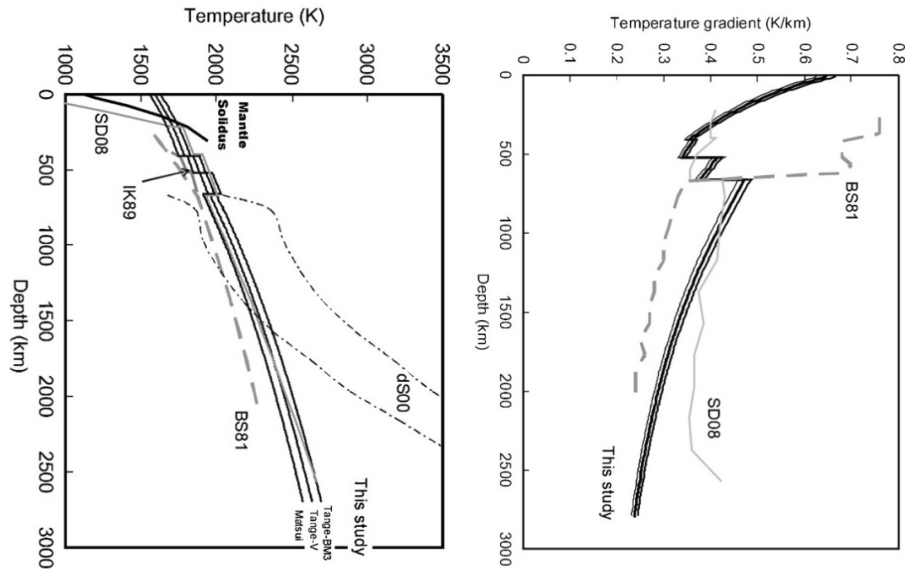
Taken from Katsura *et al.* (2010) [717]. Left: The adiabatic temperature distributions in the mantle. The three solid lines denote the temperature distributions proposed in this study using three different pressure scales. Those proposed by the previous studies are shown for comparison (BS81: Brown and Shankland, 1981; IK89: Ito and Katsura, 1989; dS00: da Silva *et al.*, 2000; SD08: Stacey and Davis, 2008). The mantle solidus proposed by Hirschmann (2000) is also shown. Right: Adiabatic temperature gradient in the mantle. The adiabatic temperature gradient abruptly increases in association with the olivine-wadsleyite, wadsleyite-ringwoodite, ringwoodite-perovskite + periclase transitions, as is the case for the thermal expansion. The adiabatic temperature gradients given in the previous studies are also shown for comparison (BS81: Brown and Shankland, 1981; SD08: Stacey and Davis, 2008 [1271]).
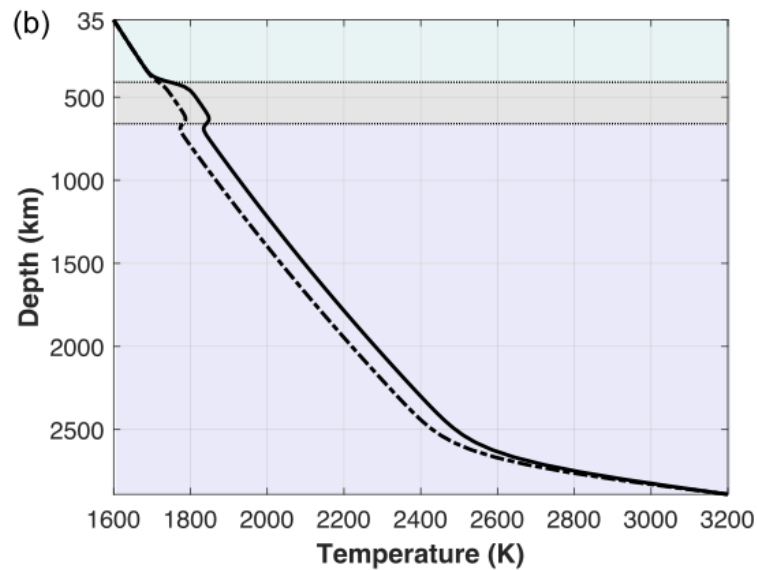
Taken from Neuharth and Mittelstaedt [991].

Prof. Steinberger was gracious to communicate to me the data of Steinberger & Calderwood (2006) [1284]:

1204
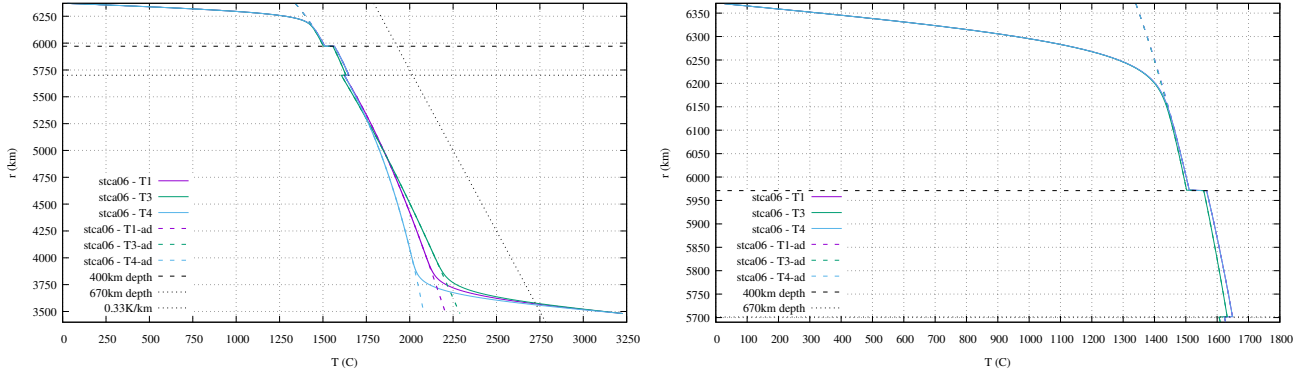
Temperature profiles from Steinberger & Calderwood [1284].

One can gather data from various papers and books and this yields the following figure[1]:
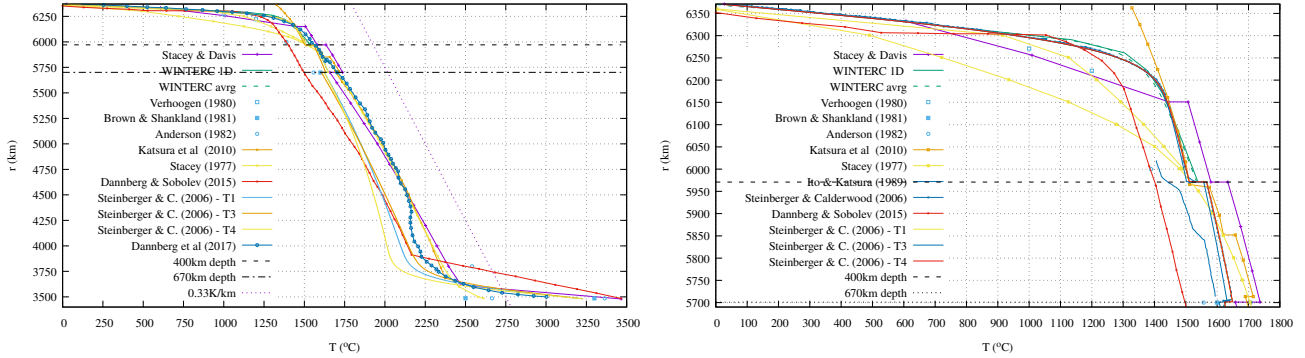
Note that the T values of Stacey (1977) [1270] inexplicably go to 540-550K at the surface. The value 545 has therefore been subtracted from the values in the paper. These then align remarquably well with those of Katsura *et al.*

We see that there is a remarkable agreement between many studies for the temperature down to 400km depth. Most show a step in the profile between 400 and 670km depth but the temperature at 670km depth seems to be between 1650°C and 1750°C. Further down, the discrepancy between studies increases (effectively boiling down to the value of the adiabatic gradient). 100km above the CMB all studies seem to indicate a temperature of 2300-2400°C. Temperature values at the CMB differ a lot and as noted in Bull *et al.* (2009) [181]: "Estimates of CMB temperature vary greatly. [...] We impose a surface temperature of 273 K and investigate three CMB temperatures: 3000 K (consistent with previous studies of this nature, e.g., Kellogg *et al.* ., 1999 [733]), 3950 K (consistent with recent results on the double-crossing of Post- Perovskite, e.g., Hernlund *et al.* ., 2005 [598]; van der Hilst *et al.* , 2007 [605]; Alfè *et al.* ., 2002), and 4800 K (Knittle and Jeanloz, 1991). The 3950 K temperature is the 'reference' temperature which we use for the majority of cases." The CMB temperature is set to $2900 - 3200$°C in Fowler [434].

Also, oceanic plates being thinner than continental plates (alongside with different radiogenic decay and thermal properties) one can draw two representative mantle geotherms [1378]:

---

[1]It must be said that finding actual data -not figures- is remarkably difficult.
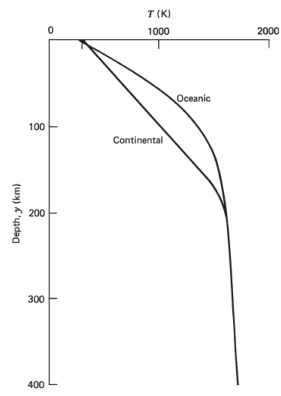
Figure 4.56 Representative oceanic and continental shallow upper mantle geotherms.

Taken from Turcotte & Schubert [1378]

2024: Taken from Plimmer, Davies, and Panton [1068] (2024).

1206
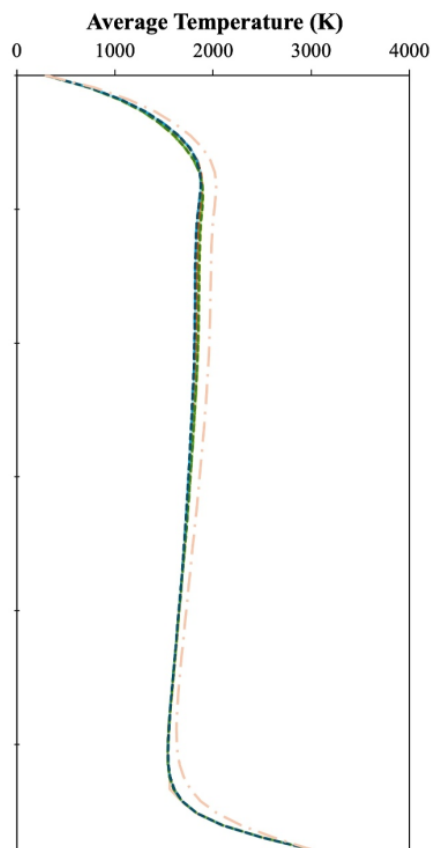
Isentropic gradient The material properties also define the slope of the adiabat (the change in temperature with pressure at constant entropy) at all pressures and temperatures. Using the cyclic relation, we can define this slope in terms of partial differentials of the entropy with respect to pressure and temperature:

$$\left(\frac{\partial T}{\partial p}\right)_S = -\left(\frac{\partial T}{\partial S}\right)_p \left(\frac{\partial S}{\partial p}\right)_T \tag{26.11}$$

$$= -\left(\frac{T}{C_p}\right)\left(-\frac{\alpha}{\rho}\right) \tag{26.12}$$

$$= \frac{\alpha T}{\rho C_p} \tag{26.13}$$

This expression does not pose a constraint on the material properties, but in order to be self-consistent, the adiabat must be computed following this relation.

For complex material models, obtaining analytical functions which obey all these relations may be a non-trivial exercise. Furthermore, it is often not immediately clear when a given formulation is thermodynamically inconsistent. Indeed, both the thermodynamic and the geodynamic literature contain many equations of state and material parameterizations which do not obey these relations! This may not invalidate the results obtained with these models, but it is a point worth keeping in mind as the geodynamics community moves to more complicated and more realistic parameterizations.

A final note of warning: Some compressible formulations in ASPECT *(Section* **??***) use the isothermal compressibility, while others use the isentropic compressibility. Fully self-consistent material models must either specify what approximation of the compressible equations they are consistent with (see Section* **??***), or have a switch so that they use the correct compressibility for each of the different approximations. The conversion between isothermal and isentropic compressibilities is given in (***??***).*

Initial conditions and the adiabatic pressure/temperature The thermo-mechanically coupled (Navier-)Stokes equations require us to pose initial conditions for the temperature Note that the equations themselves do not require that initial conditions are specified for the velocity and pressure variables (since there are no time derivatives on these variables in the model).

Nevertheless, a nonlinear solver will have difficulty converging to the correct solution if we start with a completely unphysical pressure for models in which coefficients such as density ρ and viscosity η depend on the pressure and temperature. To this end, ASPECT uses pressure and temperature fields $p_{\text{ad}}(z), T_{\text{ad}}(z)$ computed in the adiabatic conditions model (see Section **??**). By default, these fields satisfy adiabatic conditions:

$$\rho C_p \frac{\mathrm{d}}{\mathrm{d}z} T_{\text{ad}}(z) = \frac{\partial \rho}{\partial T} T_{\text{ad}}(z) g_z, \tag{26.14}$$

$$\frac{\mathrm{d}}{\mathrm{d}z} p_{\text{ad}}(z) = \rho g_z, \tag{26.15}$$

where strictly speaking g_z is the magnitude of the vertical component of the gravity vector field, but in practice we take the magnitude of the entire gravity vector.

These equations can be integrated numerically starting at $z = 0$, using the depth dependent gravity field and values of the coefficients $\rho = \rho(p, T, z), C_p = C_p(p, T, z)$. As starting conditions at $z = 0$ we choose a pressure $p_{\text{ad}}(0)$ equal to the average surface pressure (often chosen to be zero, see

Section **??**), and an adiabatic surface temperature $T_{\text{ad}}(0)$ that is also selected in the input parameter file.

> **Note:** The adiabatic surface temperature is often chosen significantly higher than the actual surface temperature. For example, on earth, the actual surface temperature is on the order of 290 K, whereas a reasonable adiabatic surface temperature is maybe 1600 K. The reason is that the bulk of the mantle is more or less in thermal equilibrium with a thermal profile that corresponds to the latter temperature, whereas the very low actual surface temperature and the very high bottom temperature at the core-mantle boundary simply induce a thermal boundary layer. Since the temperature and pressure profile we compute using the equations above are simply meant to be good starting points for nonlinear solvers, it is important to choose this profile in such a way that it covers most of the mantle well; choosing an adiabatic surface temperature of 290 K would yield a temperature and pressure profile that is wrong almost throughout the entire mantle.

For instance, let us consider $\alpha = 3 \cdot 10^{-5}$, $g_z = 10$, $C_p = 1250$, $\rho = \rho_0(1 - \alpha(T - T_0)$ so that $\frac{\partial \rho}{\partial T} = -\alpha \rho_0$ with $\rho_0 = 3300$.

Then we must solve the following equation

$$\rho_0(1 - \alpha(T - T_0))C_p \frac{\mathrm{d}T}{\mathrm{d}z} = -\alpha T^2 g_z$$

———————————-

In Verhoogen (1951) [1418]: As is well known, the adiabatic gradient may be written as

$$\frac{dT}{dP} = \alpha T / \rho C_p$$

If hydrostatic equilibrium is assumed, the pressure varies with depth h as $dP = \rho g dh$, so that

$$\frac{d \ln T}{dh} = \frac{\alpha g}{C_p}$$

from which the temperature T at any depth h may be computed as a function of the temperature at any assigned depth if the ratio α/C_p is known at all depths (g, the acceleration of gravity, will be taken as constant in the mantle).

———————————-

———————————-

From DyMaLi: In the interior of a convecting medium temperatures follow an adiabatic profile. At the top and bottom of a convecting layer thermal boundary layers with large thermal gradients form. The interior is thermally well mixed and therefore essentially isothermal, with a slight increase of temperatures with depth due to the effect of pressure. For example in the Earth's mantle the geothermal gradient $\partial T/\partial z$ is about 20C/km near the surface and about 0.3C/km in the interior of the mantle. This small gradient in the interior is the adiabatic gradient. If a small volume of material is moved to shallower depth is experiences a slight increase in volume due to the decreasing pressure and associated with this a slight decrease in temperature. This change in temperature is the adiabatic temperature change.

The adiabatic gradient can be determined from the thermodynamics relation between entropy per unit mass S, temperature T, and pressure P:

$$dS = \left(\frac{dS}{dT}\right)_P dT + \left(\frac{dS}{dP}\right)_T dP = \frac{C_p}{T} dT - \frac{\alpha}{\rho} dP$$

In case of a reversible adiabatic process the entropy change is zero, and so the adiabatic gradient is:

$$\left(\frac{dT}{dP}\right)_S = \frac{\alpha T}{\rho C_p}$$

The gradient can also be expressed in terms of depth, remembering that $dp = \rho g dz$ in a hydrostatic fluid:

$$\left(\frac{dT}{dz}\right)_S = \frac{\alpha g T}{C_p}$$

Thus to determine the adiabatic gradient one needs values of α and C_p with depth. These are obtained from laboratory experiments.

One also needs an estimate of density as a function of depth, which is generally determined from seismology. Integration of the adiabatic gradient in terms of pressure then gives temperature as a function of pressure. Temperature as a function of depth is obtained by integrating the density distribution to obtain g as a function of depth.

not finished ——————————————-

Vol07_02

This adiabaticity hypothesis should, however, not be taken too literally (Jeanloz and Morris, 1987 [679]). In most numerical simulations, the resulting averaged geotherm can be far (a few hundred kelvins) from adiabatic (Bunge *et al.* , 2001 [185]). First, radioactive heating, dissipation, and diffusion are never totally negligible, second, even if each fluid parcel follows its own adiabatic geotherm, the average geotherm may not correspond to any particular adiabat.

————————————————- From Steinberger oct 26th, 2020:

Yes, I think the temperature below the lithosphere is still fairly well-constrained, based on magmas produced on mid-oceanic ridges (away from plumes) if one takes these as representative. Regarding Dannberg and Sobolev (not Solomatov!) I don't know why it is lower. In their supplementary figures, the extrapolation to the surface is actually ~ 1250K, not 1250°C, which is even less. Perhaps best to directly ask Juliane about this.

How the temperature increases with depth is more uncertain; the temperature gradient is often taken as adiabatic; that's what I also assumed in the 2006 paper, then it is defined by a ordinary differential equation (eq. 12 in that paper). I think the main uncertainty is the thermal expansivity. In my model, it strongly decreases with depth, I guess that is why my models have a lower temperature in the lower mantle than other models. I think that strong decrease in expansivity is still the consensus, but I didn't closely follow the literature recently. But the temperature gradient between the thermal boundary layers may actually be subadiabatic, hence even lower, due to cold slabs sinking to and accumulating above the CMB, and hot plume material feeding into the asthenosphere, below the lithosphere. So, yes, I would say the difference in the deep mantle (if not more) reflects the current state in the community.

And I think the uncertainites of CMB temperature are even larger. And I would of course be happy if you host my data on a public github repo. Just for the viscosity profile, there should be some explanation given that it shouldn't necessarily be taken "at face value" but that (like in the 2006 paper) it is possible to multiply different parts of the profile with different "scaling viscosities".

————————————————-

🛡 Relevant Literature: *On the thermal gradient in the Earth's deep interior*, Tirone (2016) [1359]
Is the mantle geotherm subadiabatic, Jeanloz & Morris (1987) [679]
Mantle convection, the asthenosphere, and Earth's thermal history King (2015) [746]

check section 7.7 in Fowler !
boundary layer jape82
evolution shpe79

26.5 Earth: radial thermal expansion profile
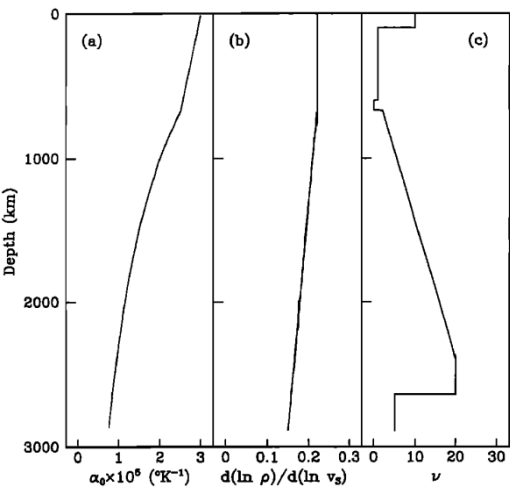
Taken from Pari and Peltier [1033] (1995).
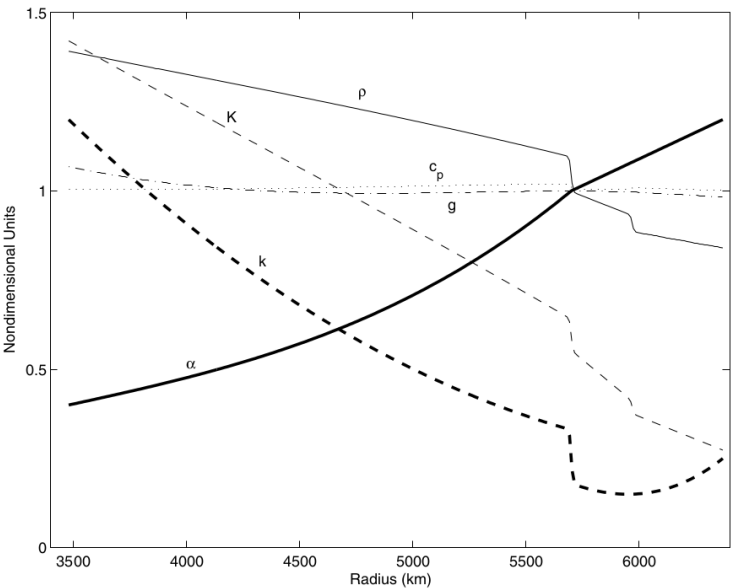
Taken from Butler and Jarvis [210] (2004). To dimensionalize, multiply the thermal expansivity (thick solid line) by $\alpha_0 = 2 \cdot 10^{-5}$ K^{-1}.
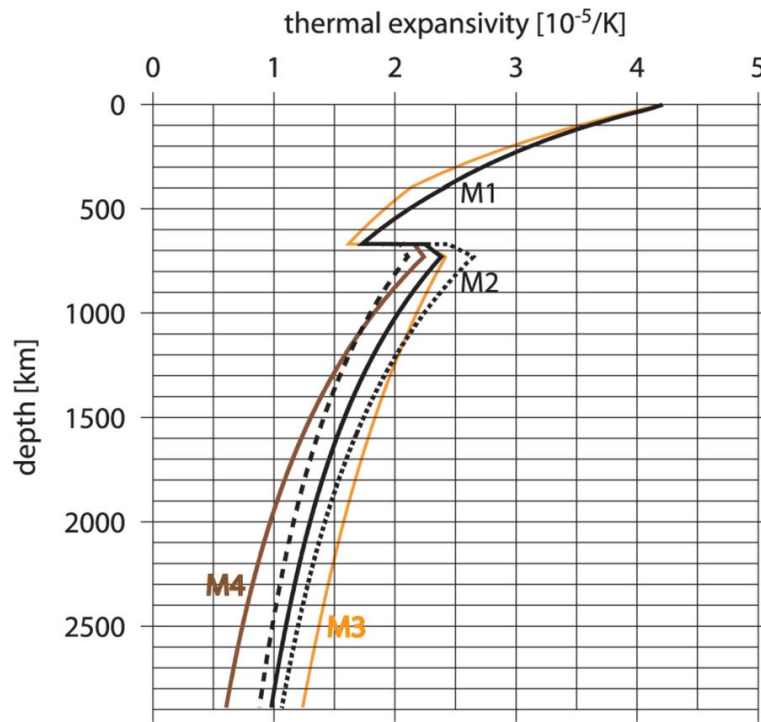
Taken from Steinberger and Calderwood [1284].

see Matyska *et al.* (2011) [893]

Mantle convection with internal heating and pressure-dependent thermal expansivity, Leitch *et al.* (1991) [809]

Eq(8) of Hassan *et al.* [586]
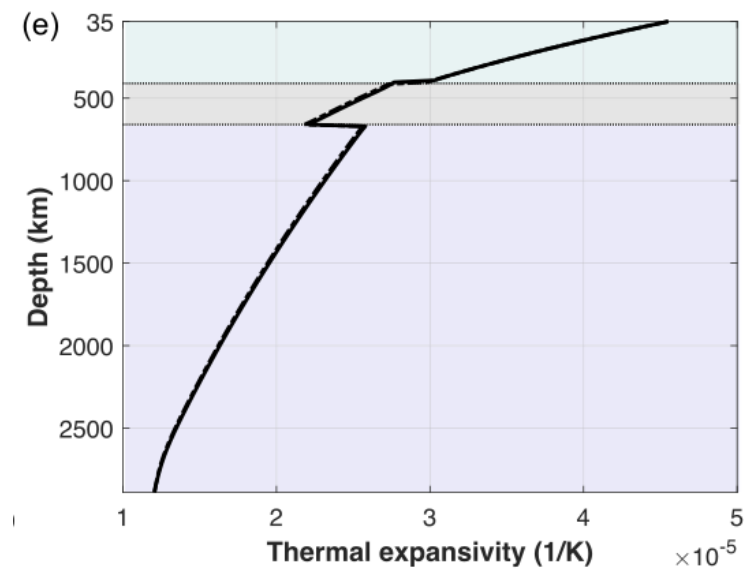
Taken from Neuharth and Mittelstaedt [991].

26.6 Earth: radial density profile
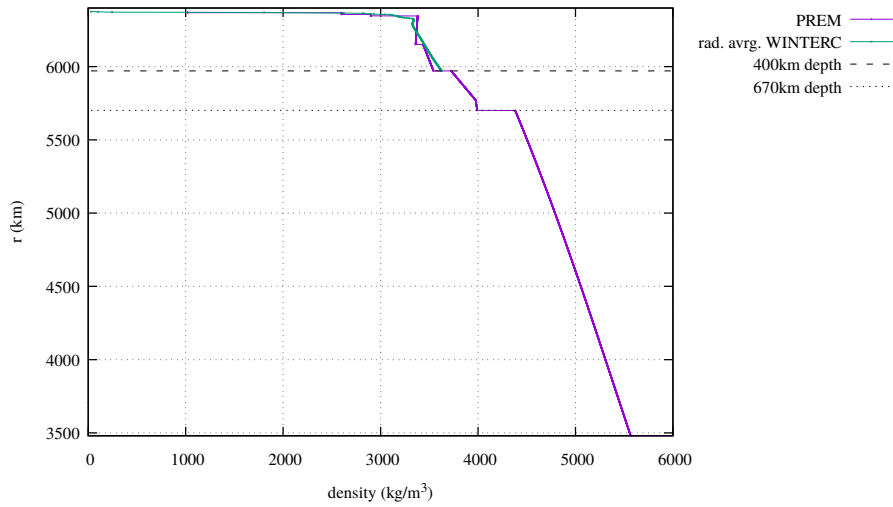
🛢 Relevant Literature: Kennett (1998) [735]
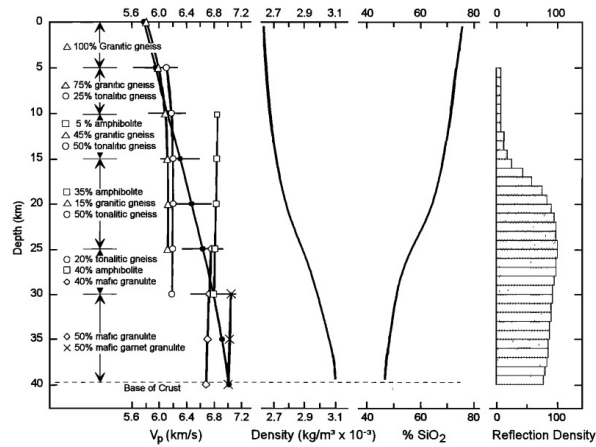
Taken from Christensen and Mooney [258] (1995). A model for average crustal petrology versus depth consistent with average velocity depth profile (solid circles) and velocity depth curves for common rock types (open symbols). Variations of density and SiO 2 content with depth are from rock percentages shown on the left.

Let us look at the density and pressure profiles in a 1D isothermal 'planet'. We start from

$$-\vec{\nabla}p + \rho\vec{g} = \vec{0}$$

In 1D, and assuming $\vec{g} = -g\vec{e}_z$:

$$-\frac{dp}{dz} - \rho g = 0$$

or

$$\frac{dp}{dz} = -\rho g$$

Assuming ρ and g constant in the domain $z \in [0, L]$, we can solve this ODE and we obtain:

$$p(z) = \rho g(L - z)$$

1212

Let us now turn to the case of an isothermal but compressible fluid. Its density is now given by

$$\rho(p) = \rho_0(1 - \beta p)$$

where β is the compressibility (assumed to be constant in the domain). We must then solve

$$
\frac{dp}{dz} = -\rho_0(1 - \beta p)g \;\; \Rightarrow \;\; \frac{dp}{1 - \beta p} = -\rho_0 g dz
$$
$$
\Rightarrow \;\; \int \frac{dp}{1 - \beta p} = -\int \rho_0 g dz
$$
$$
\Rightarrow \;\; -\frac{1}{\beta}\ln(1 - \beta p) = -\rho_0 g z + C
$$
$$
\Rightarrow \;\; \ln(1 - \beta p) = \beta\rho_0 g z + D
$$
$$
\Rightarrow \;\; 1 - \beta p = \exp\left(\beta\rho_0 g z + D\right)
$$
$$
\Rightarrow \;\; p(z) = \frac{1}{\beta}\left[1 - \exp\left(\beta\rho_0 g z + D\right)\right]
$$

(26.16)

At $z = L$ we require $p = 0$ so we obtain

$$p(z) = \frac{1}{\beta}\left[1 - \exp\left(\beta\rho_0 g(z - L)\right)\right]$$

Note that when the compressibility tends to zero, by virtue of

$$\exp x \sim 1 + x + \frac{x^2}{2} + ...$$

for $x \to 0$ we then recover the linear pressure profile above.

Let us now take $\rho_0 = 4000 \, \mathrm{kg\,m^{-3}}$, $g = 10 \, \mathrm{m\,s^{-2}}$ and $\beta = 4 \cdot 10^{-12} \, \mathrm{Pa^{-1}}$ [465] and $L = 3000$ km.
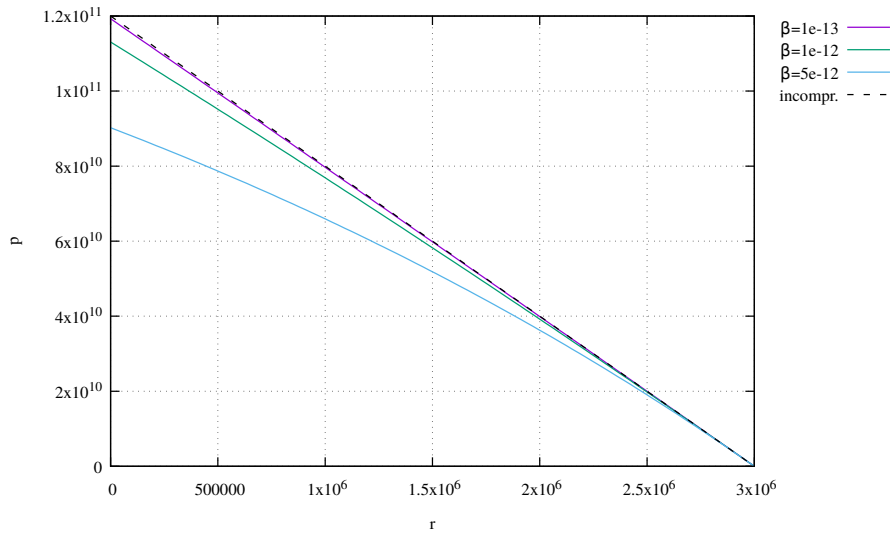
TODO: produce same plot with density
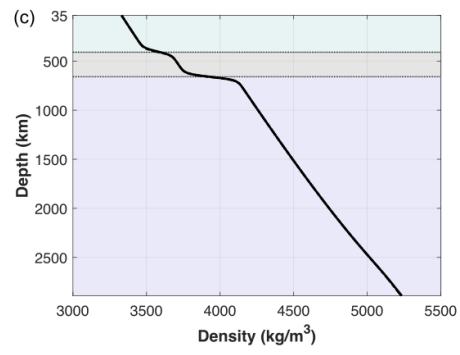
Taken from Neuharth and Mittelstaedt [991].

26.7 Earth: radial thermal conductivity profile
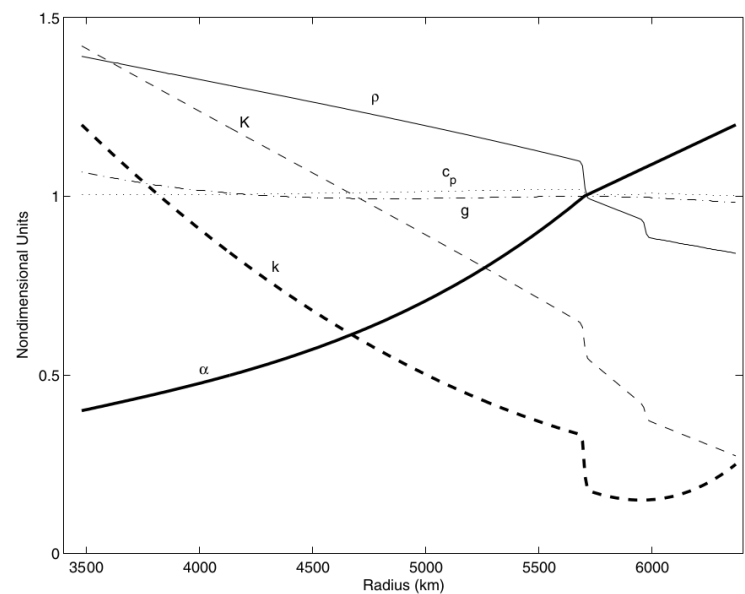
Taken from Butler and Jarvis [210] (2004). To dimensionalize, multiply the thermal conductivity (thick dashed line) by $k_0 = 10\ \mathrm{W\,m^{-1}\,K^{-1}}$.
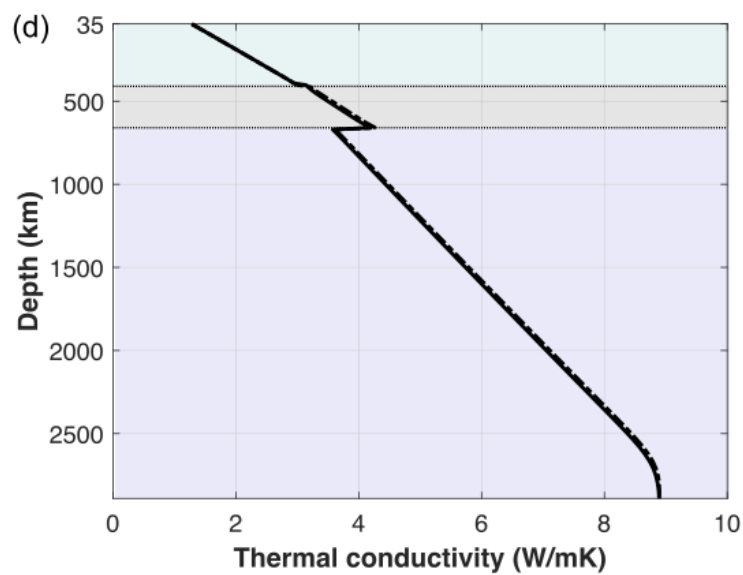
Taken from Neuharth and Mittelstaedt [991].

26.8 Earth: radial heat capacity profile
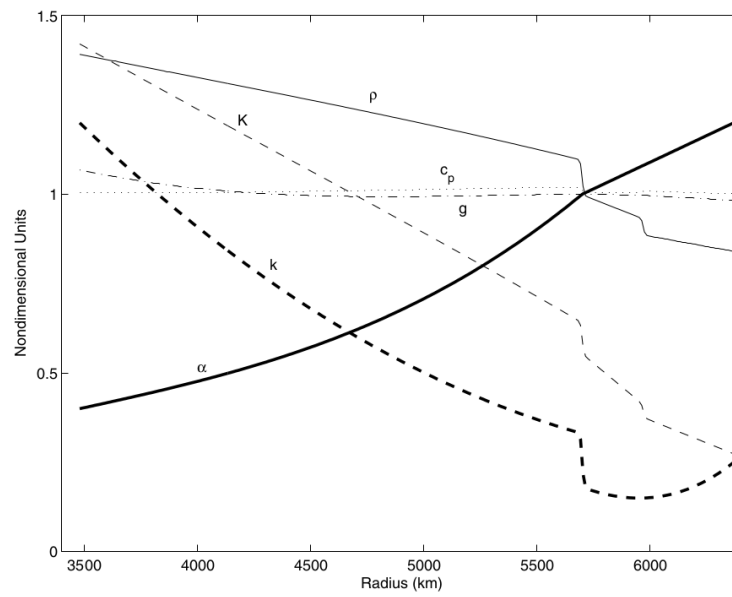
Taken from Butler and Jarvis [210] (2004). To dimensionalize, multiply the specific heat capacity (dotted line) by $c_{p,0} = 1250 \ \mathrm{J\,kg^{-1}\,K^{-1}}$.

26.9 Earth: Crust and lithosphere structure

- Crust 5.1: Mooney, Laske, and Masters [947] (1998).
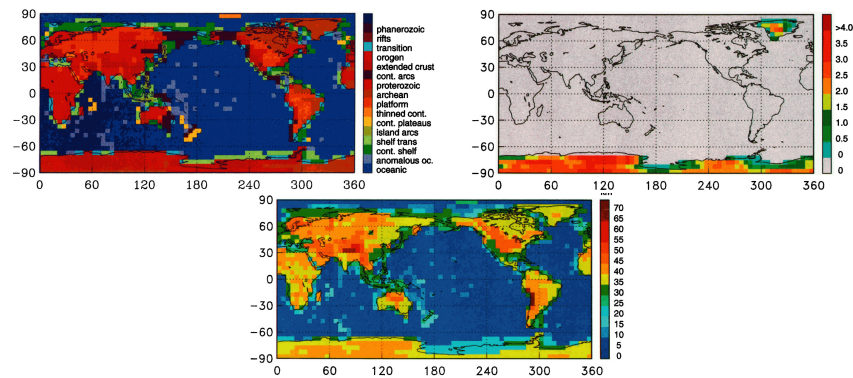
Taken from Mooney, Laske, and Masters [947].

- Crust 1.0: `https://igppweb.ucsd.edu/~gabi/crust1.html`

- Litho 1.0: `https://igppweb.ucsd.edu/~gabi/litho1.0.html`

26.10 Mars structure

- James M Dohm, Shigenori Maruyama, Motoyuki Kido, and Victor R Baker. "A possible anorthositic continent of early Mars and the role of planetary size for the inception of Earth-like life". In: *Geoscience Frontiers* 9.4 (2018), pp. 1085–1098. DOI: 10.1016/j.gsf.2016.12.003

  Mars' radius (˜3390 km) is about half that of the Earth, and it has nearly 1/8 of Earth's mass. Though a lack of seismic information means that its interior structure and composition are not well known as Earth's, Mars is, nevertheless, inferred to be a differentiated planetary body. It has a core estimated to be nearly the same radius as Earth's solid inner core (˜1200 km); a thinner mantle (˜2100 km), about ˜2/3 that of Earth (˜2100 km); and a thicker crust, nearly double or more than that of Earth, reaching more than 60 km thick in parts of the southern cratered highlands. Oceanic-type crust has been interpreted to be roughly representative of the northern plains, while continental crust has been hypothesized to underlie the southern highlands
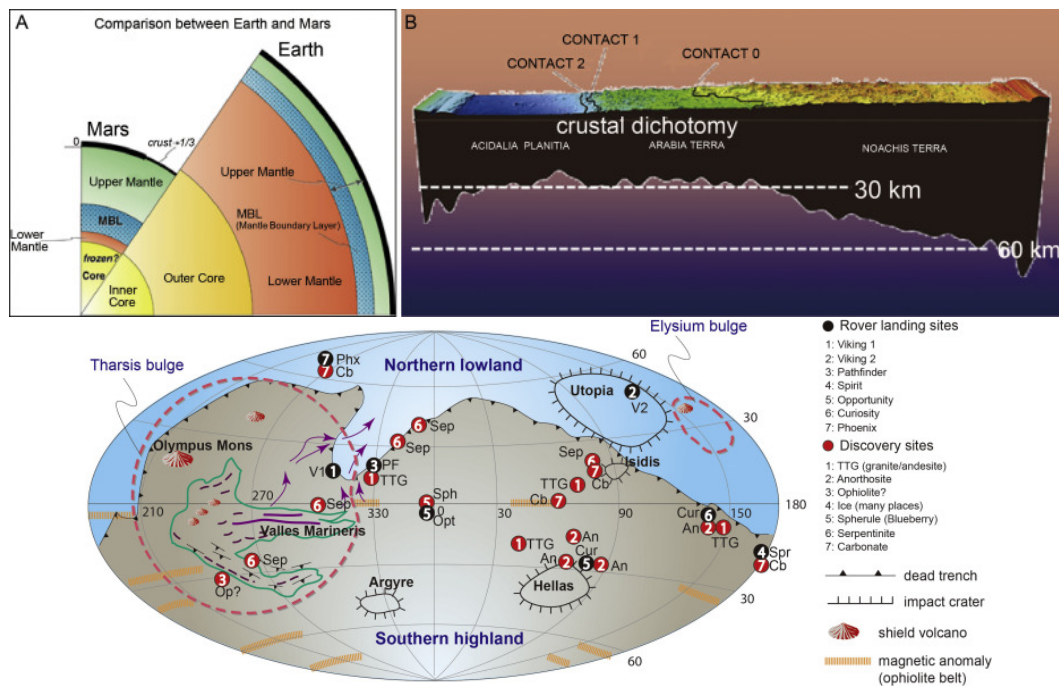
- A-C Plesa et al. "The thermal state and interior structure of Mars". In: *Geophys. Res. Lett.* 45.22 (2018), pp. 12–198. DOI: 10.1029/2018GL080728
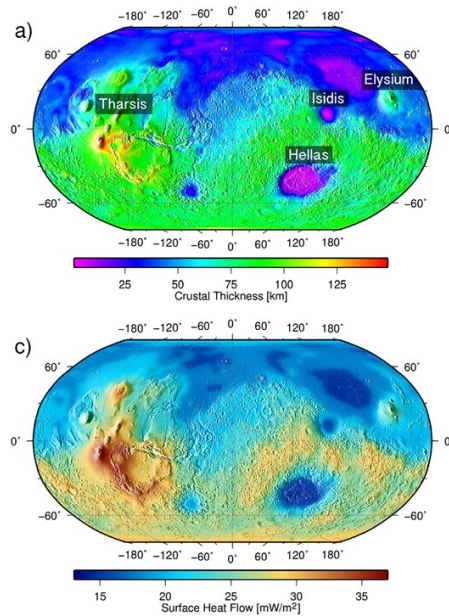
- Min Ding, Jian Lin, Chen Gu, Qinghua Huang, and Maria T Zuber. "Variations in Martian lithospheric strength based on gravity/topography analysis". In: *J. Geophys. Res.: Planets* 124.11 (2019), pp. 3095–3118. DOI: 10.1029/2019JE005937
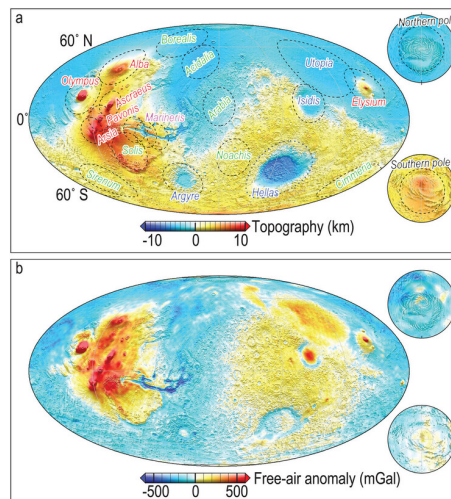
- Suzanne E Smrekar et al. "Pre-mission InSights on the interior of Mars". In: *Space Science Reviews* 215.1 (2019), pp. 1–72. DOI: 10.1007/s11214-018-0563-9
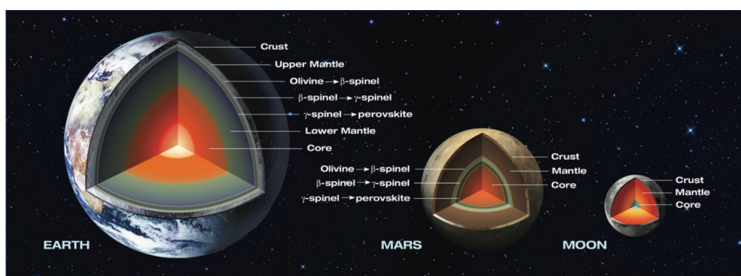
Fig. 4 Representative crustal thickness model (top) using the interior density profile for the model DWTh2Ref1. The crustal densities for the southern highlands and northern lowlands (bottom) are 2900 and 3000 kg m^{-3}, respectively. The minimum crustal thickness was constrained to be 1 km, which determines the average crustal thickness to be 42 km and the crustal thickness at the InSight landing site (star) to be 32 km. Data are presented in Mollweide projections centered over the Tharsis province (100° W), and grid lines are spaced every 30 in latitude and longitude. The dictomomy boundary used in the lower image is taken from Andrews-Hanna et al. (2008)
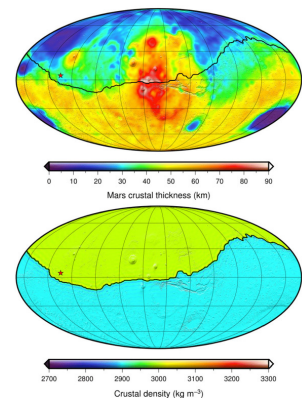
- Brigitte Knapmeyer-Endrun et al. "Thickness and structure of the martian crust from InSight seismic data". In: *Science* 373.6553 (2021), pp. 438–443. DOI: 10.1126/science.abf8966

  we determine the structure of the crust beneath the InSight landing site on Mars using both marsquake recordings and the ambient wavefield. By analyzing seismic phases that are reflected and converted at subsurface interfaces, we find that the observations are consistent with models with at least two and possibly three interfaces. If the second interface is the boundary of the crust, the thickness is 20 ± 5 kilometers, whereas if the third interface is the boundary, the thickness is 39 ± 8 kilometers. Global maps of gravity and topography allow extrapolation of this point measurement to the whole planet, showing that the average thickness of the martian crust lies between 24 and 72 kilometers. Independent bulk composition and geodynamic constraints show that the thicker model is consistent with the abundances of crustal heat-producing elements observed for the shallow surface, whereas the thinner model requires greater concentration at depth.
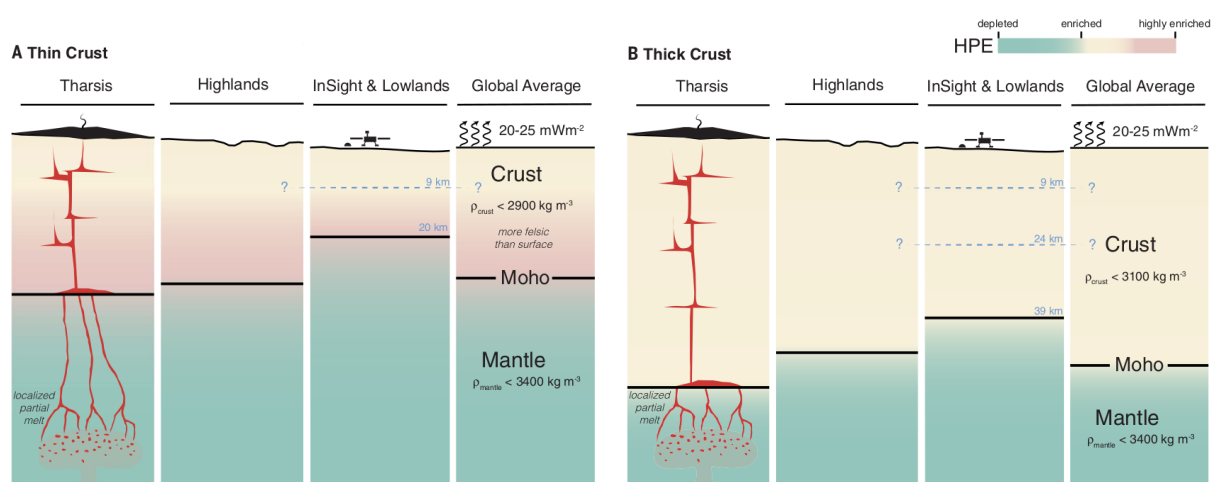
- Simon C. Stähler et al. "Seismic detection of the martian core". In: *Science* 373.6553 (2021), pp. 443–448. DOI: 10.1126/science.abi7730

  We detected reflections of seismic waves from the core-mantle boundary of Mars using InSight seismic data and inverted these together with geodetic data to constrain the radius of the liquid metal core to 1830 ± 40 kilometers. The large core implies a martian mantle mineralogically similar to the terrestrial upper mantle and transition zone but differing from Earth by not having a bridgmanite-dominated lower mantle. We inferred a mean core density of 5.7 to 6.3 grams per cubic centimeter, which requires a substantial complement of light elements dissolved in the iron-nickel core.

InSight

Cerberus
Fossae

Core Radius:
1830±40 km
fluid, no S-waves

Inner
Core?

Composition:
Fe-S + light elements

Density:
5.8-6.2 g/cm³

Core
Shadow

Tharsis

S-waves:
*S,SKS,**ScS***

P-waves:
P, PKP
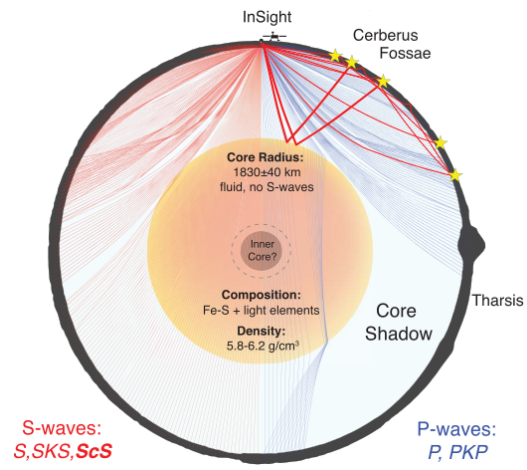
26.11 Mars: radial viscosity profile

Please note that the figures below are screen captures and omit to specify how these results were arrived at.

- J. H. Roberts and S. Zhong. "Degree-1 convection in the Martian mantle and the origin of the hemispheric dichotomy". In: *J. Geophys. Res.: Solid Earth* 111.E6 (2006). DOI: 10.1029/2005JE002668
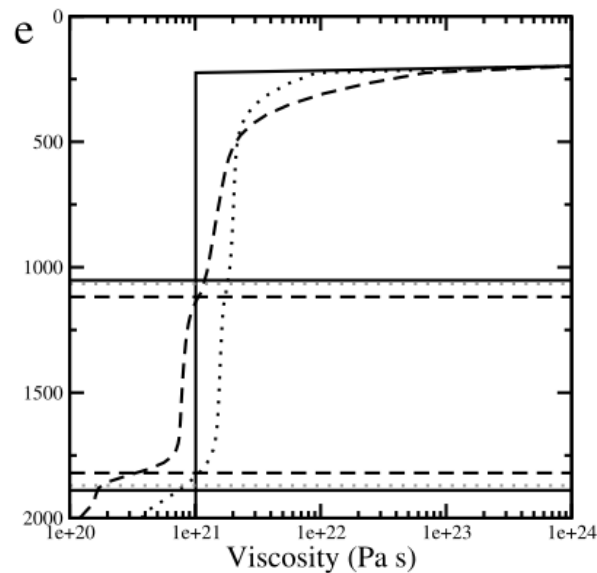
- Tobias Keller and Paul J Tackley. "Towards self-consistent modeling of the martian dichotomy: The influence of one-ridge convection on crustal thickness distribution". In: *Icarus* 202.2 (2009), pp. 429–443. DOI: 10.1016/j.icarus.2009.03.029

- Bernhard Steinberger, Stephanie C Werner, and Trond H Torsvik. "Deep versus shallow origin of gravity anomalies, topography and volcanism on Earth, Venus and Mars". In: *Icarus* 207.2 (2010), pp. 564–577. DOI: 10.1016/j.icarus.2009.12.025

1222
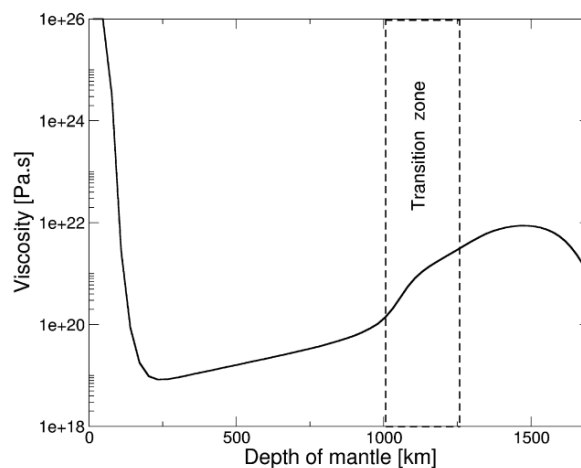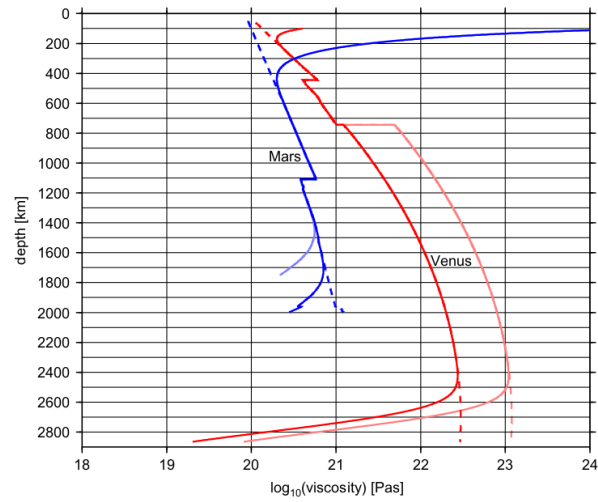
- J. H. Roberts and J. Arkani-Hamed. "Impact-induced mantle dynamics on Mars". In: *Icarus* 218.1 (2012), pp. 278–289. DOI: 10.1016/j.icarus.2011.11.038
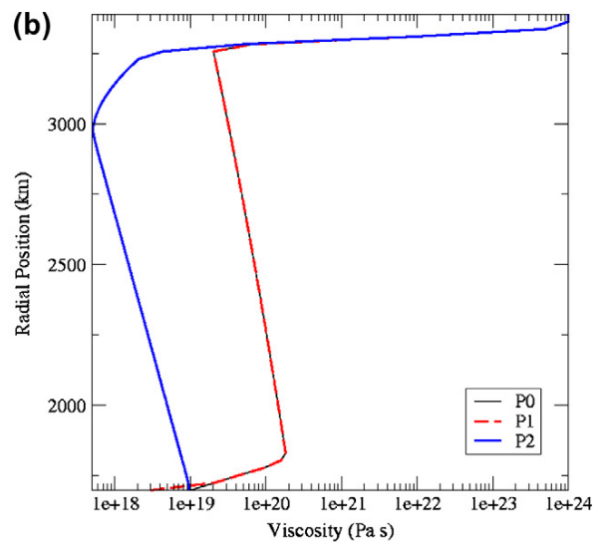
- Thomas Ruedas, Paul J. Tackley, and Sean C. Solomon. "Thermal and compositional evolution of the martian mantle: Effects of phase transitions and melting". In: *Phys. Earth. Planet. Inter.* 216 (2013), pp. 32–58. DOI: 10.1016/j.pepi.2012.12.002
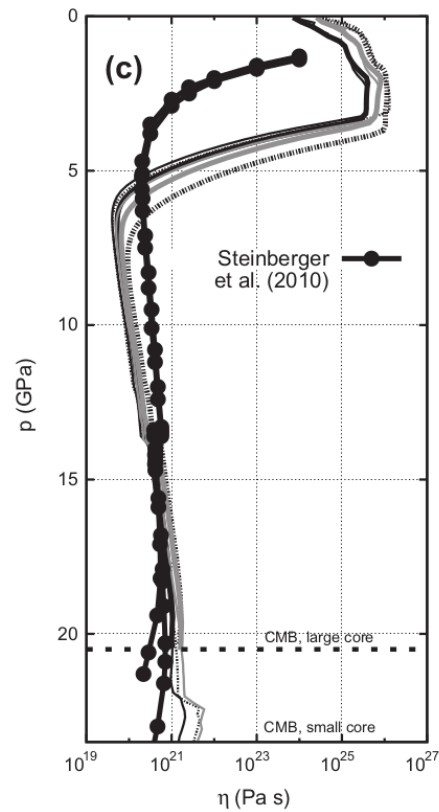
- P. Sekhar and S.D. King. "3D spherical models of Martian mantle convection constrained by melting history". In: *Earth Planet. Sci. Lett.* 388 (2014), pp. 27–37. DOI: `10.1016/j.epsl.2013.11.047`
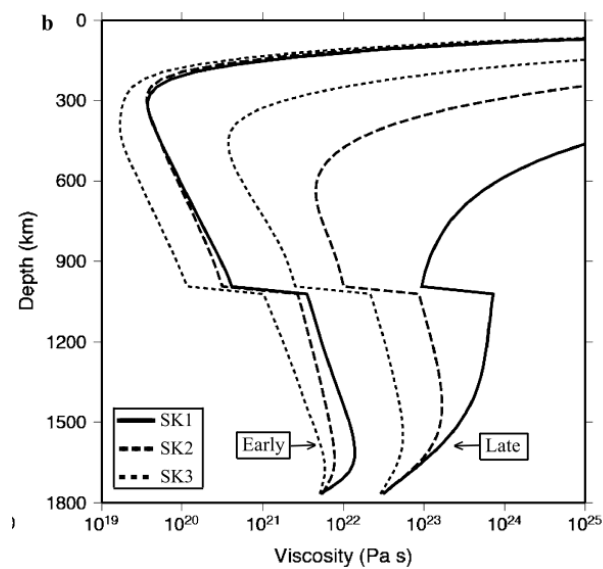
- Amir Khan et al. "A geophysical perspective on the bulk composition of Mars". In: *J. Geophys. Res.: Planets* 123.2 (2018), pp. 575–611. DOI: `10.1002/2017JE005371`
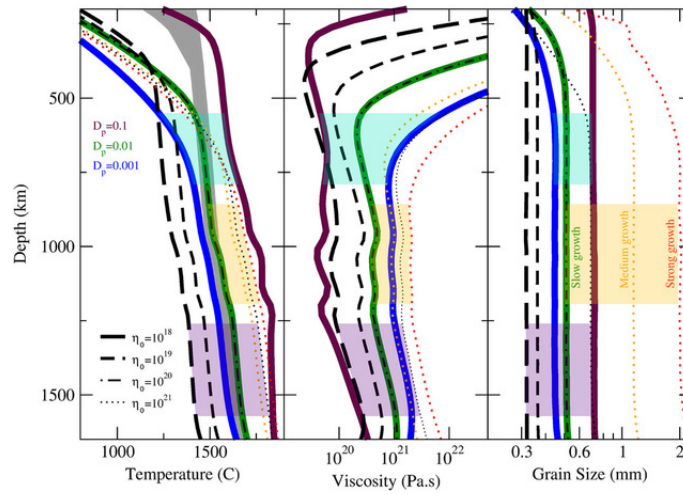
- A-C Plesa et al. "The thermal state and interior structure of Mars". In: *Geophys. Res. Lett.* 45.22 (2018), pp. 12–198. DOI: 10.1029/2018GL080728
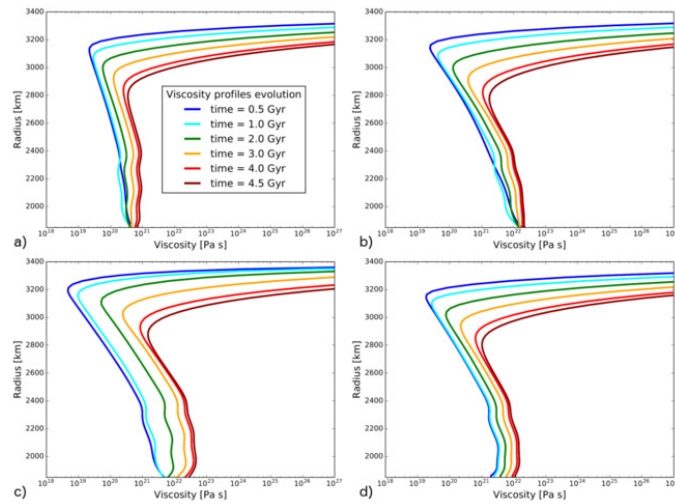
see supplementary material for legend.

- Ana-Catalina Plesa, Mark Wieczorek, Martin Knapmeyer, Attilio Rivoldini, Michaela Walterova, and Doris Breuer. "Interior dynamics and thermal evolution of Mars–a geodynamic perspective". In: *Advances in geophysics*. Vol. 63. 2022, pp. 179–230. DOI: 10.1016/bs.agph.2022.07.005
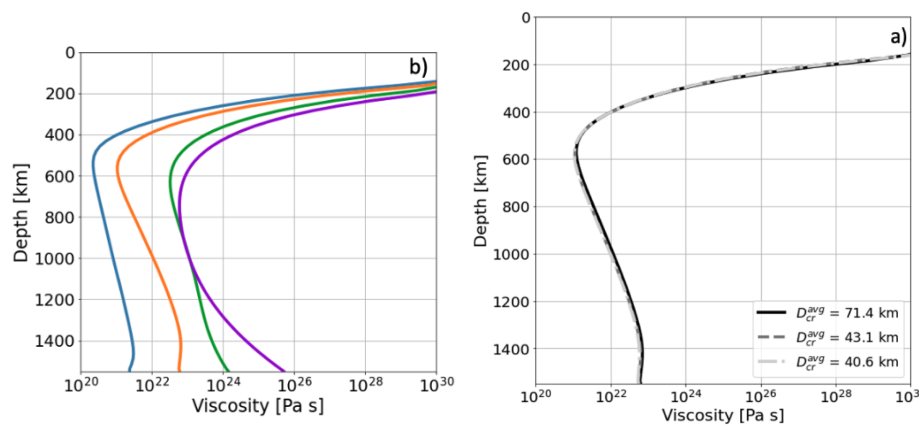
26.12 Mars: radial temperature profile

Please note that the figures below are screen captures and omit to specify how these results were arrived at.

- Doris Breuer, Dave A Yuen, and Tilman Spohn. "Phase transitions in the Martian mantle: Implications for partially layered convection". In: *Earth Planet. Sci. Lett.* 148.3-4 (1997), pp. 457–469. DOI: 10.1016/S0012-821X(97)00049-6
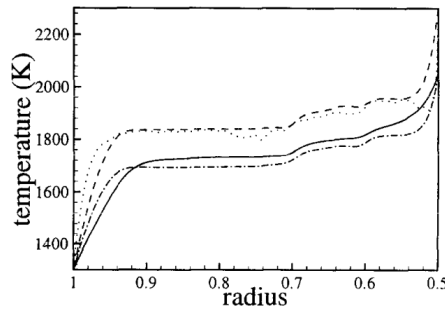
- J. H. Roberts and S. Zhong. "Degree-1 convection in the Martian mantle and the origin of the hemispheric dichotomy". In: *J. Geophys. Res.: Solid Earth* 111.E6 (2006). DOI: 10.1029/2005JE002668
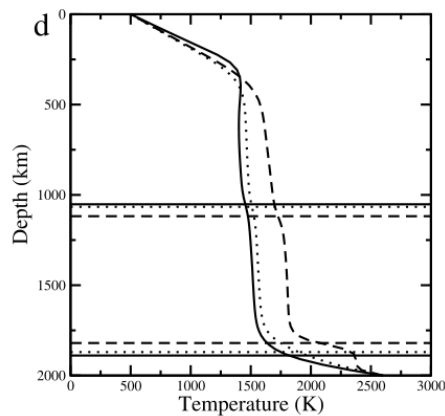
- Bernhard Steinberger, Stephanie C Werner, and Trond H Torsvik. "Deep versus shallow origin of gravity anomalies, topography and volcanism on Earth, Venus and Mars". In: *Icarus* 207.2 (2010), pp. 564–577. DOI: 10.1016/j.icarus.2009.12.025
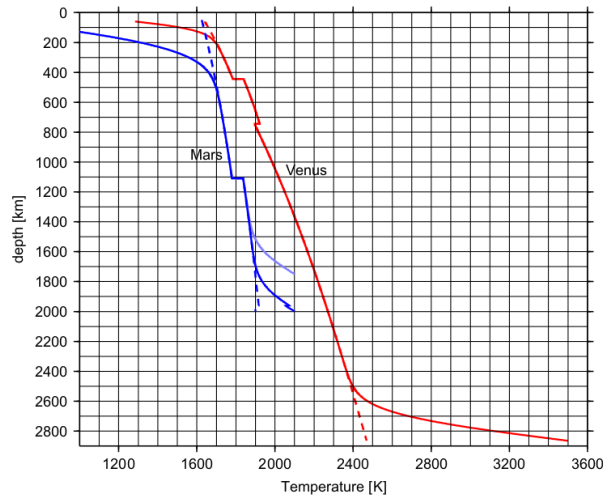
- J. H. Roberts and J. Arkani-Hamed. "Impact-induced mantle dynamics on Mars". In: *Icarus* 218.1 (2012), pp. 278–289. DOI: 10.1016/j.icarus.2011.11.038
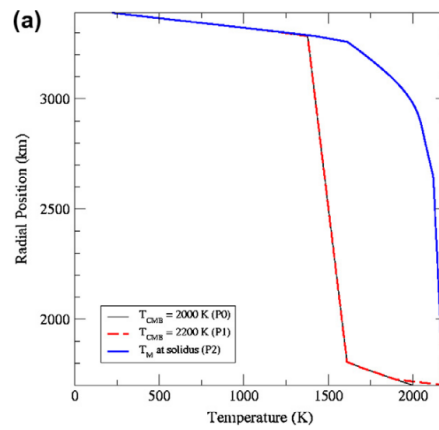
- S. Azuma and I. Katayama. "Evolution of the rheological structure of Mars". In: *Earth, Planets and Space* 69.1 (2017), pp. 1–13. DOI: 10.1186/s40623-016-0593-z

- Amir Khan et al. "A geophysical perspective on the bulk composition of Mars". In: *J. Geophys. Res.: Planets* 123.2 (2018), pp. 575–611. DOI: 10.1002/2017JE005371

1227
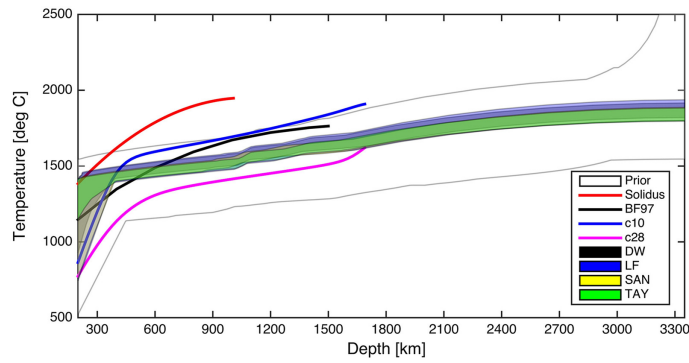
- A-C Plesa et al. "The thermal state and interior structure of Mars". In: *Geophys. Res. Lett.* 45.22 (2018), pp. 12–198. DOI: 10.1029/2018GL080728

- Suzanne E Smrekar et al. "Pre-mission InSights on the interior of Mars". In: *Space Science Reviews* 215.1 (2019), pp. 1–72. DOI: 10.1007/s11214-018-0563-9
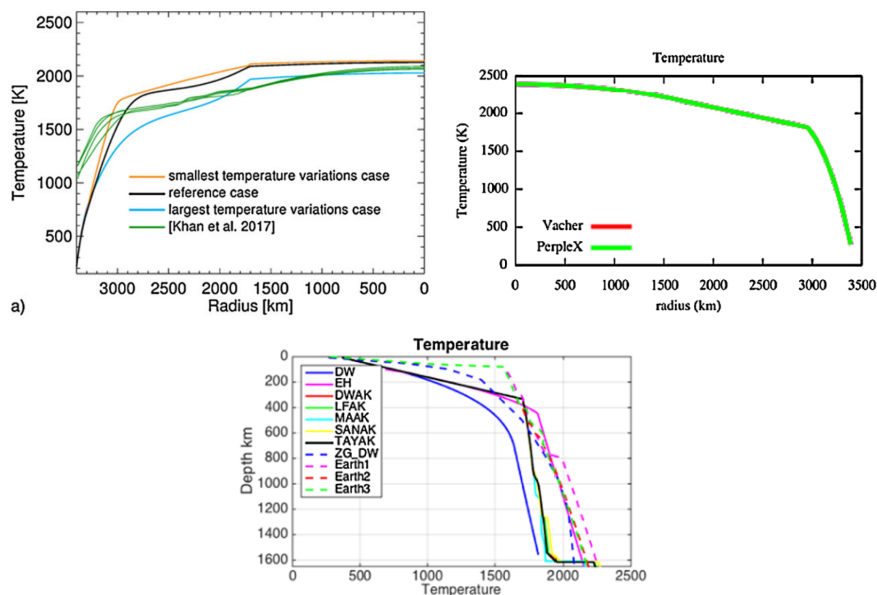
- Henri Samuel, Maxim D Ballmer, Sebastiano Padovan, Nicola Tosi, Attilio Rivoldini, and Ana-catalina Plesa. "The Thermo-Chemical Evolution of Mars With a Strongly Stratified Mantle". In: *J. Geophys. Res.: Planets* 126.4 (2021), e2020JE006613. DOI: 10.1029/2020JE006613

- Ana-Catalina Plesa, Mark Wieczorek, Martin Knapmeyer, Attilio Rivoldini, Michaela Walterova, and Doris Breuer. "Interior dynamics and thermal evolution of Mars–a geodynamic perspective". In: *Advances in geophysics*. Vol. 63. 2022, pp. 179–230. DOI: 10.1016/bs.agph.2022.07.005

- Henri Samuel et al. "Geophysical evidence for an enriched molten silicate layer above Mars's core". In: *Nature* 622.7984 (2023), pp. 712–717. DOI: 10.1038/s41586-023-06601-8

Left: Homogeneous mantle (no basal mantle layer); Right: Heterogeneous mantle (with a basal mantle layer)
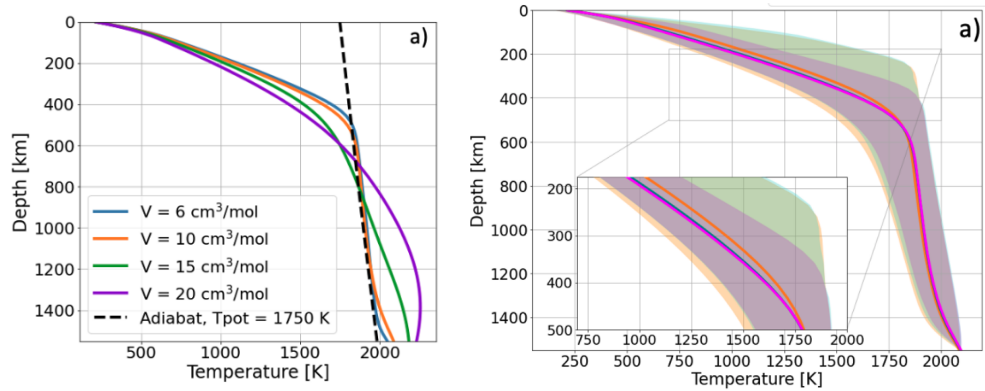
- Amir Khan, Dongyang Huang, C Durán, Paolo Angelo Sossi, Domenico Giardini, and Motohiko Murakami. "Evidence for a liquid silicate layer atop the Martian core". In: *Nature* 622.7984 (2023), pp. 718–723. DOI: `10.1038/s41586-023-06586-4`

- Mélanie Drilleau et al. "Constraints on lateral variations of Martian crustal thickness from seismological and gravity field measurements". In: *Geophys. Res. Lett.* 51.4 (2024), e2023GL105701. DOI: `10.1029/2023GL105701`
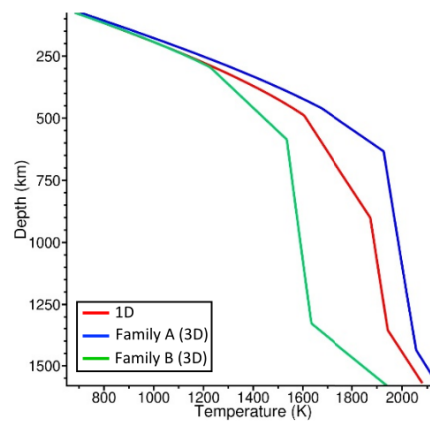
26.13 Mars: radial density profile

- Helmut Harder. "Phase transitions and the three-dimensional planform of thermal convection in the Martian mantle". In: *J. Geophys. Res.: Planets* 103.E7 (1998), pp. 16775–16797. DOI: 10.1029/98JE01543
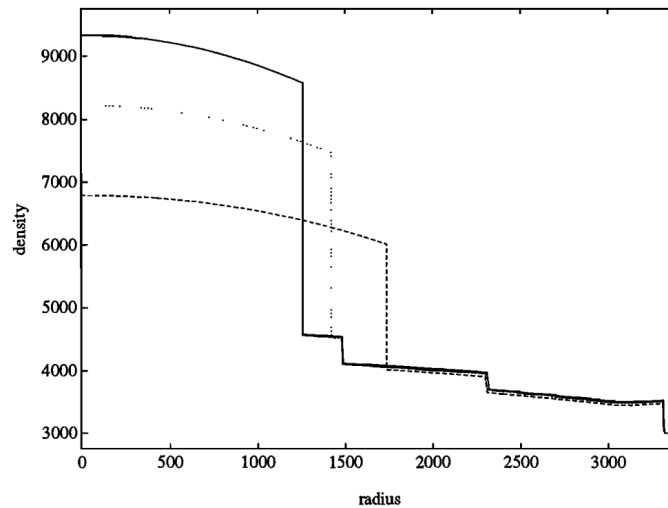
Possible density structures of Mars' solid(0%), dotted(14%), and dashed (36.5%) lines indicate sulphur content of the core.

- Constance M Bertka and Yingwei Fei. "Density profile of an SNC model Martian interior and the moment-of-inertia factor of Mars". In: *Earth Planet. Sci. Lett.* 157.1-2 (1998), pp. 79–88. DOI: 10.1016/S0012-821X(98)00030-2
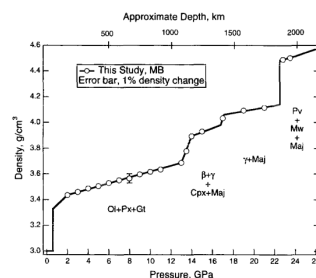
Fig. 1. Density profile of the Dreibus and Wänke [7] model Martian mantle. Mantle density calculated from the experimental results of [11] are shown with circles. The profile assumes a 50-km, 3.0-g/cm³ crust.
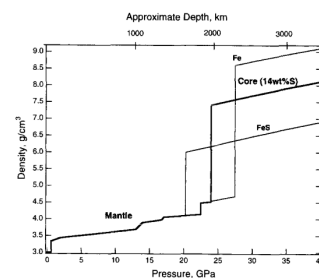
Fig. 2. Density profile of the Dreibus and Wänke [7] model Martian mantle and density profiles for a range of model core compositions, Fe to FeS. The mantle profile assumes a 50-km, 3.0-g/cm³ crust.

- Tobias Keller and Paul J Tackley. "Towards self-consistent modeling of the martian dichotomy: The influence of one-ridge convection on crustal thickness distribution". In: *Icarus* 202.2 (2009), pp. 429–443. DOI: 10.1016/j.icarus.2009.03.029
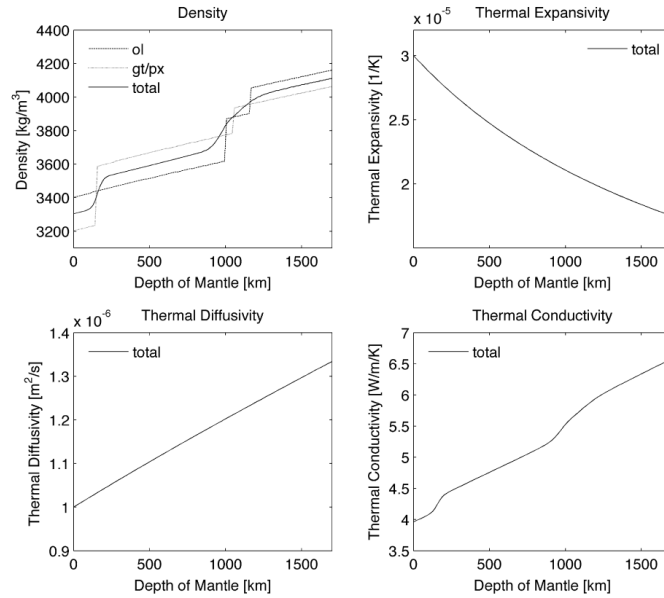
- Thomas Ruedas, Paul J. Tackley, and Sean C. Solomon. "Thermal and compositional evolution of the martian mantle: Effects of phase transitions and melting". In: *Phys. Earth. Planet. Inter.* 216 (2013), pp. 32–58. DOI: 10.1016/j.pepi.2012.12.002
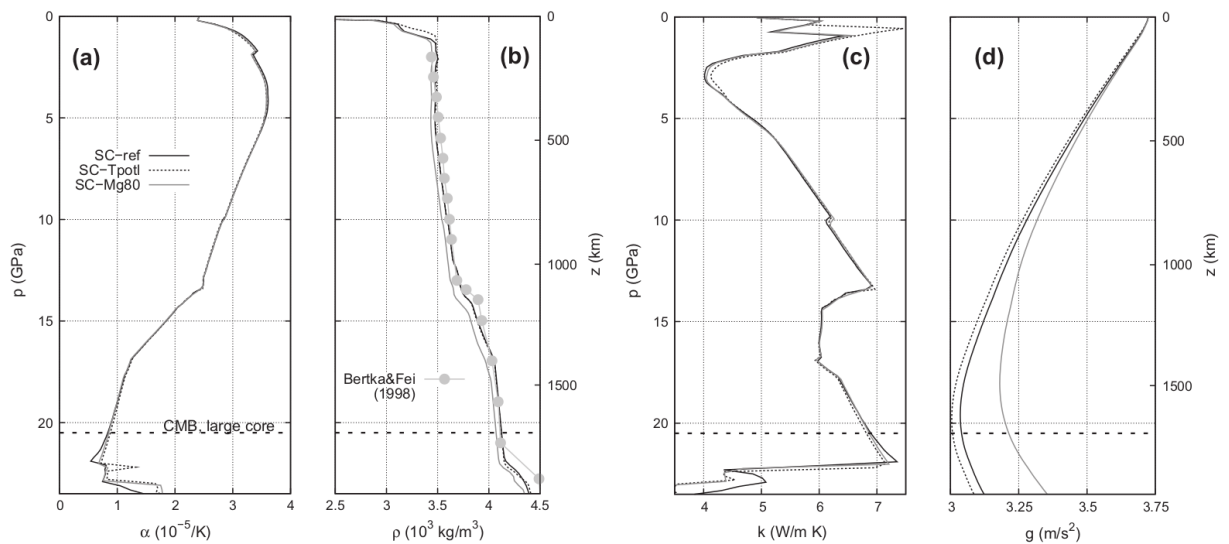
- Amir Khan et al. "A geophysical perspective on the bulk composition of Mars". In: *J. Geophys. Res.: Planets* 123.2 (2018), pp. 575–611. DOI: 10.1002/2017JE005371
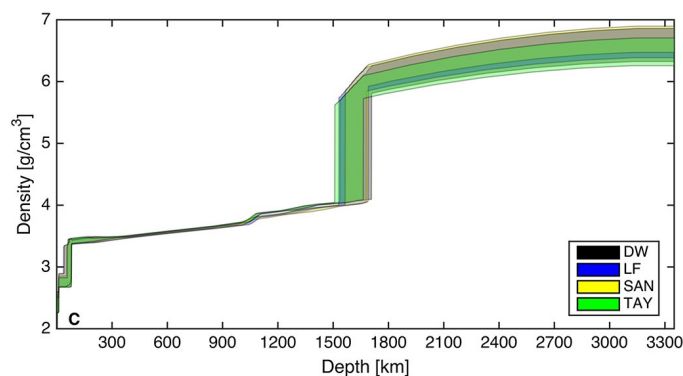
- Suzanne E Smrekar et al. "Pre-mission InSights on the interior of Mars". In: *Space Science Reviews* 215.1 (2019), pp. 1–72. DOI: 10.1007/s11214-018-0563-9
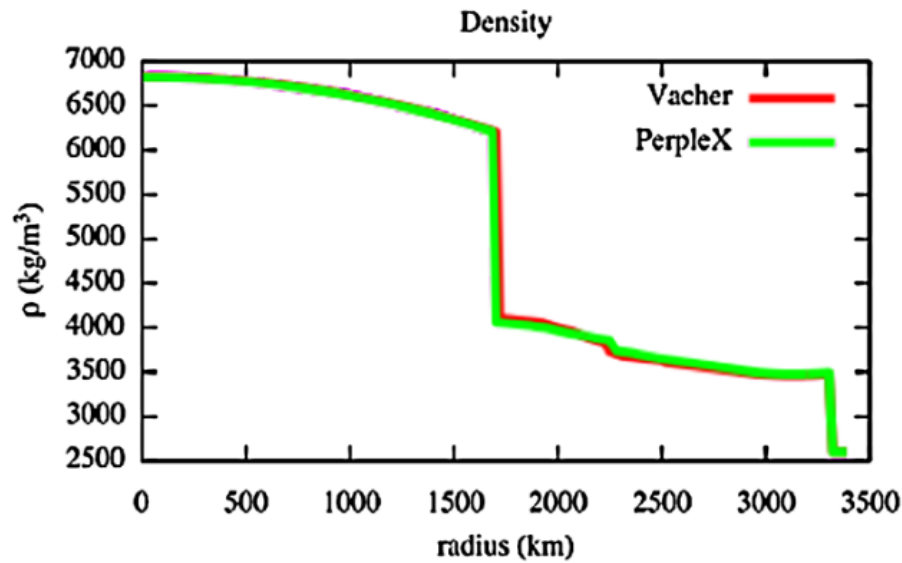
- Matthew C Brennan, Rebecca A Fischer, and Jessica CE Irving. "Core formation and geophysical properties of Mars". In: *Earth Planet. Sci. Lett.* 530 (2020), p. 115923. DOI: 10.1016/j.epsl.2019.115923

- Ana-Catalina Plesa, Mark Wieczorek, Martin Knapmeyer, Attilio Rivoldini, Michaela Walterova, and Doris Breuer. "Interior dynamics and thermal evolution of Mars–a geodynamic perspective". In: *Advances in geophysics.* Vol. 63. 2022, pp. 179–230. DOI: 10.1016/bs.agph.2022.07.005
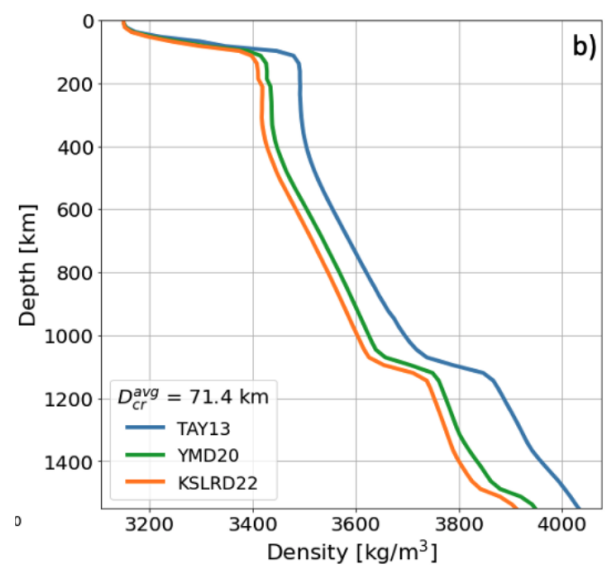
- Henri Samuel et al. "Geophysical evidence for an enriched molten silicate layer above Mars's core". In: *Nature* 622.7984 (2023), pp. 712–717. DOI: 10.1038/s41586-023-06601-8

Left: Homogeneous mantle (no basal mantle layer); Right: Heterogeneous mantle (with a basal mantle layer)
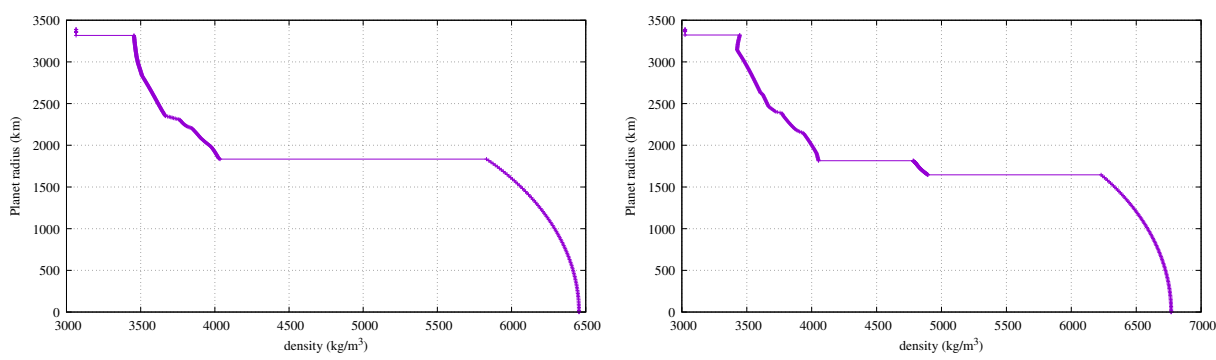
Raw data. Left: homogeneous; Right: heterogeneous.

- Amir Khan, Dongyang Huang, C Durán, Paolo Angelo Sossi, Domenico Giardini, and Motohiko Murakami. "Evidence for a liquid silicate layer atop the Martian core". In: *Nature* 622.7984 (2023), pp. 718–723. DOI: 10.1038/s41586-023-06586-4
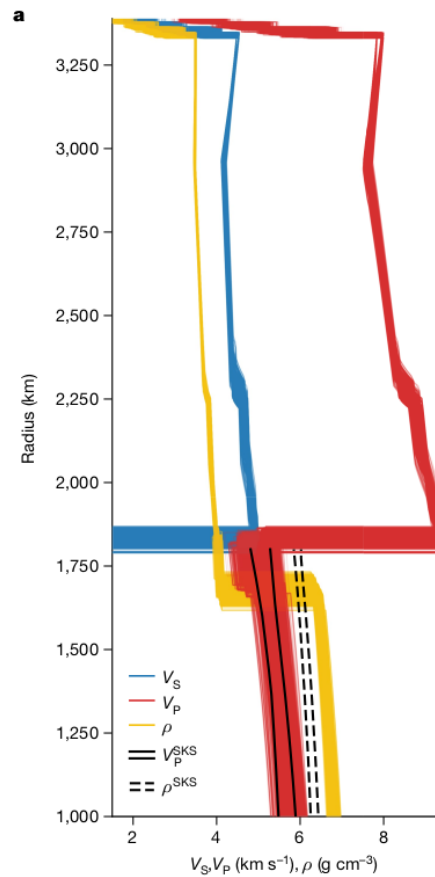
There are 1000 profiles in the online data at https://dataverse.ipgp.fr/file.xhtml?fileId=3694&version=2.0. I have selected about 20 of them to plot:
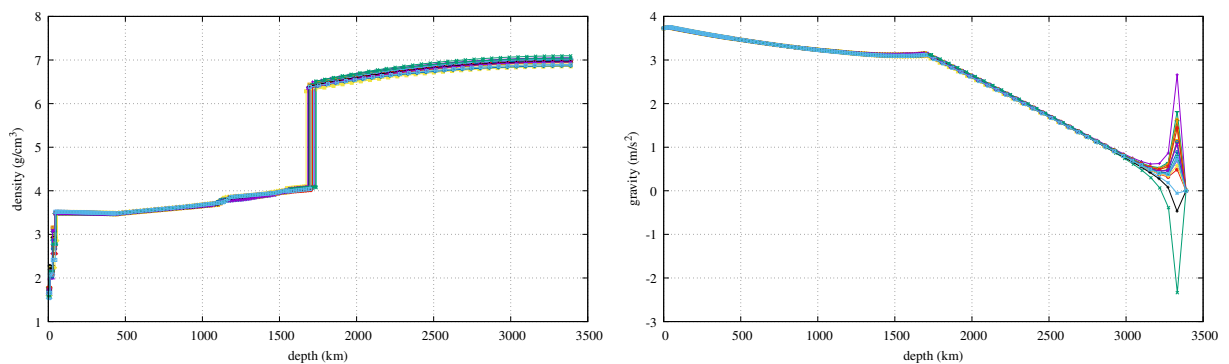
- Alex Guinard, Agnes Fienga, Anthony Memin, and Clement Ganino. "Coupled geophysical and thermal constraints link between Mars basal molten layer and the planet viscosity profile". In: *arXiv preprint arXiv:2403.06763* (2024)
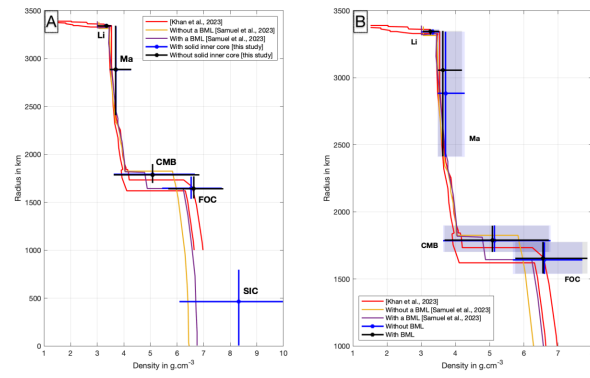
Chapter 27

[WIP] The Propagator matrix method (PMM) in geodynamics

```
giba66 <--- haoc81(appendix)  <--- riha84  <-- hacr85
            haoc78?                  riff84
            haoc79?                  hage84
```

27.1   Literature

- Freeman Gilbert and George E Backus. "Propagator matrices in elastic wave and vibration problems". In: *Geophysics* 31.2 (1966), pp. 326–332. DOI: 10.1190/1.1439771

  An early article about the PMM, often cited. Not so useful since it is not concerned with the Stokes equations of mantle dynamics.

- B.H. Hager and R.J. O'Connell. "Subduction zone dip angles and flow driven by plate motion". In: *Tectonophysics* 50 (1978), pp. 111–133. DOI: 10.1016/0040-1951(78)90130-0

  3d formulation based on spherical harmonics.

  > Kinematic models of the large scale flow in the mantle accompanying the observed plate motions are calculated by neglecting thermal buoyancy forces. The large scale flow is therefore determined by the mass flux imposed by the moving plates. The energy and momentum equations decouple, and with the assumption of a radially symmetric Newtonian viscosity, the flow accompanying the plate motions can be obtained using harmonic analysis and propagator matrices. [...] We also assume that viscosity is a function of radius only. This assumption may well neglect some important effects since the mantle temperature distribution is not radially symmetric, and viscosity is highly dependent upon temperature. The flow is assumed to be incompressible.

- Bradford H Hager and Richard J O'Connell. "Kinematic models of large-scale flow in the Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 84.B3 (1979), pp. 1031–1048. DOI: 10.1029/JB084iB03p01031

  Very similar in scope and technique to their 1978 paper.

- B.H. Hager and R.J. O'Connell. "A simple global model of plate dynamics and mantle convection". In: *J. Geophys. Res.: Solid Earth* 86.B6 (1981), pp. 4843–4867. DOI: 10.1029/JB086iB06p04843

  This article is cited a lot when people refer to the PMM. It presents in its appendix the PMM for both 2d and 3d cases.

  > The procedure for solving these equations [mass & momentum conservation equations] (givenin the appendix)is to separate the vertical and horizontal dependences of the flow parameters. The total variation of these parameters, assumed periodic with horizontal period L, is expressed in terms of a Fourier series, with each term containing a coefficient varying with depth multiplied by a harmonic function of the horizontal coordinate.The flow equations can be reduced to a coupled set of first-order differential equations in depth. These are solved analytically by using the propagator matrix technique [492].

- Mark A Richards and Bradford H Hager. "Geoid anomalies in a dynamic Earth". In: *J. Geophys. Res.: Solid Earth* 89.B7 (1984), pp. 5987–6002

  Refers to Hager and O'Connell [549] (1981).

  > A familiar and useful property of the propagator matrix formulation is that solution vectors can be propagated through a seriesof differentmaterial layers by simply forming the product of the individual layer matrices [...]. Therefore changes in viscosity (and density) with depth are easily incorporated into this formalism.

- Y. Ricard, L. Fleitout, and C. Froidevaux. "Geoid heights and lithospheric stresses for a dynamic Earth". In: *Annales Geophysicae* 2.3 (1984), pp. 267–286. DOI: xxxx

  3d, refers to Hager and O'Connell [550] (1978).

- Bradford H Hager. "Subducted slabs and the geoid: Constraints on mantle rheology and flow". In: *J. Geophys. Res.: Solid Earth* 89.B7 (1984), pp. 6003–6015. DOI: 10.1029/JB089iB07p06003

- Bradford H Hager, Robert W Clayton, Mark A Richards, Robert P Comer, and Adam M Dziewonski. "Lower mantle heterogeneity, dynamic topography and the geoid". In: *Nature* 313.6003 (1985), p. 541. DOI: 10.1038/313541a0

- Neil M Ribe. *Theoretical Mantle Dynamics*. Cambridge University Press, 2018

  Ribe cites Gantmacher[1] as all others, but actually provides more information: vol I, p 120, Eq. (53). The two pages in the book rely heavily on [549]. We find the same content in [1130].

- Shijie Zhong. "Analytic solutions for Stokes' flow with lateral variations in viscosity". In: *Geophy. J. Int.* 124.1 (1996), pp. 18–28. DOI: 10.1111/j.1365-246X.1996.tb06349.x

  > Analytic solutions for 2-D incompressible Stokes' flow with lateral variations in viscosity have been developed with a Green's function method and matrix propagator techniques. The analytic solutions are developed based on the observation that lateral variations in viscosity only result in mode coupling between viscosity and buoyancy in the horizontal dimension and not in the vertical dimension.

---

[1] https://www.maths.ed.ac.uk/~v1ranick/papers/gantmacher1.pdf

- Zdeněk Martinec, Detlef Wolf, et al. "Explicit form of the propagator matrix for a multi-layered, incompressible viscoelastic sphere". In: (1998)

- B.A. Buffet, C.W. Gable, and R.J. R.J. O'Connell. "Linear stability of a layered fluid with mobile surface plates". In: *J. Geophys. Res.: Solid Earth* 99(B10) (1994), pp. 19, 885–19, 900

- Nicola Tosi. "Numerical modeling of present-day mantle convection". PhD thesis. Charles University in Prague, 2007

    After introducing the partial differential equations that govern the flow in the mantle and the gravitational potential, we briefly review the technique of the matrix propagator which is generally employed to solve such equations in a spherical geometry under the approximation of laterally homogeneous viscosity. The method allows us to obtain an analytical solution in terms of the spherical harmonic expansion of the field quantities. In the presence of lateral viscosity variations, this method is no longer applicable.

- W. Leng and S. Zhong. "Viscous heating, adiabatic heating and energetic consistency in compressible mantle convection". In: *Geophy. J. Int.* 173 (2008), pp. 693–702. DOI: `10.1111/j.1365-246X.2008.03745.x`

- X. Liu and S. Zhong. "Analyses of marginal stability, heat transfer and boundary layer properties for thermal convection in a compressible fluid with infinite Prandtl number". In: *Geophy. J. Int.* 194 (2013), pp. 125–144. DOI: `10.1093/gji/ggt117`

    For compressible convection with depth-dependent density and possibly other depth-dependent thermodynamic properties, propagator matrix method is more effective. The propagator matrix method has been used to obtain analytic solution of the Stokes flow problem for incompressible (Hager and O'Connell [549], 1981) and compressible (Leng and Zhong [816], 2008) models. With a stream-function and vorticity formulation, Jarvis and McKenzie [672] (1980) employed the propagator matrix method for marginal stability analysis with heat flux boundary conditions. Buffet, Gable, and R.J. O'Connell [173] used the propagator matrix method for marginal stability analysis for incompressible flows with depth-dependent viscosity. In this study, we develop a new implementation of propagator matrix method for marginal stability analysis for both incompressible and compressible flows with free-slip and isothermal boundary conditions. Our implementation is based on a stress-velocity formulation which is similar to that in Leng and Zhong [816] (2008), but we also incorporate the linearized energy equation. The setup of propagator matrix and the solution procedure are discussed in Appendix A.

- Puskar Mondal and Jun Korenaga. "A propagator matrix method for the Rayleigh–Taylor instability of multiple layers: a case study on crustal delamination in the early Earth". In: *Geophy. J. Int.* 212.3 (2018), pp. 1890–1901. DOI: `10.1093/gji/ggx513`

    In this 2-D Cartesian geometry, we expand field variables in the Fourier series assuming the horizontal invariance of material properties.

redo in a stone?

- Chuan Qin, Shijie Zhong, and Roger Phillips. "Formation of the lunar fossil bulges and its implication for the early Earth and Moon". In: *Geophys. Res. Lett.* 45.3 (2018), pp. 1286–1296. DOI: `10.1002/2017GL076278`

27.2 Theory in 2d

27.3 Theory in 3d

Chapter 28

[WIP] Numeriek modelleren van geodynamische processen

28.1 Inleiding

De zin en noodzaak van modelexperimenten met betrekking tot geodynamische processen is vooral gelegen in het volgende:

- Beperkte mogelijkheid van direkte (in situ) waarnemingen als aanvulling op indirecte bijv. seismologische- of gravimetrische waarnemingen. Denk hierbij bijvoorbeeld aan het beperkte bereik van diepboringen.

- De extreem lange tijdschaal waarop de processen plaatsvinden - postglaciale opheffing $\sim 10^5 \ yr$, mantelconvectie overturntijd $\sim 10^8 \ yr$. Het is dan praktisch onmogelijk om de afloop van de processen voldoende ver in de tijd te volgen.

Verschillende manieren van modelleren die worden toegepast zijn:

- Fysisch modelleren (laboratorium experimenten). voorbeelden van geodynamische toepassingen zijn

  - Simulatie van mantelconvectie door bestudering van Rayleigh-Benard convectie in tankmodellen.

  - Bestudering van plastische deformatie van gelaagde structuren in centrifuge experimenten.

- Theoretisch (wiskundig) modelleren. Theoretische modellen kunnen worden opgesteld voor geofysische problemen door formulering van wiskundige modelvergelijkingen. De modelvergelijkingen zijn i.h.a. partiële differentiaalvergelijkingen, die kunnen worden afgeleid uit fysische behoudswetten. Een in het vervolg veelvuldig besproken voorbeeld hiervan is de warmte (diffusie) vergelijking voor een statisch medium - zie bijv. hoofdstukken 2 en 3 - die wordt afgeleid uit een thermische energiebalans,

$$\int_V \rho c_p \frac{\partial T}{\partial t} dV = \int_{\partial V} -\mathbf{q} \cdot \mathbf{n} dA + \int_V H dV$$

waarin $\mathbf{q} \cdot \mathbf{n}$ de warmtestroomdichtheid door het oppervlak ∂V voorstelt en H de interne warmteproduktiviteit.

Dergelijke theoretische modellen hebben voor geofysische problemen een lange traditie. De gevolgde werkwijze is daarbij voornamelijk analytisch geweest. De modelvergelijkingen werden analytisch opgelost en de resulterende oplossingen - bijvoorbeeld een temperatuurveld als oplossing van de warmtevergelijking - kunnen eventueel numeriek worden gee:valueerd. Een voorbeeld hiervan vinden we bij het evalueren van de Fourier-reeks oplossing voor een afkoelende laag waarvan onder- en bovenzijde op constante temperatuur gehouden worden,

$$T(z,t) = \sum_{n=1}^{\infty} A_n sin(n\pi z/d) e^{-(n\pi/d)^2 \kappa t}$$

met

$$A_n = \frac{2}{d} \int_0^d T(z,0) sin(n\pi z/d) dz$$

Analytische oplosmethoden zijn vaak alleen geschikt voor gei:dealiseerde modellen bijvoorbeeld met vereenvoudigde geometrie van het domein of (stukwijs) constant veronderstelde fysische parameters die als coefficienten in de modelvergelijkingen optreden. Een voorbeeld van het laatste is de uniforme warmtediffusiviteit κ in bovenstaande Fourier oplossing.

In die gevallen waarin de wiskundige vergelijkingen niet analytisch kunnen worden opgelost (tot evalueerbare functies worden herleid) kan numeriek modelleren uitkomst bieden.

Kenmerkende karakteristieken van numerieke methoden van modelleren zijn:

- De oplossing wordt gezocht voor de onbekende veldgrootheden (bijv. de temperatuur in de warmtevergelijking) in een discrete verzameling punten in het oplossingsdomein (ruimte,tijd) - het rooster. Dit wordt aangeduid als de domeindiscretisatie.

- De continue modelvergelijkingen worden m.b.v. discretisatie methoden omgezet in stelsels algebraische vergelijkingen in de onbekende roosterpunt waarden.

- De algebraische vergelijkingen worden numeriek opgelost voor gegeven model parameters zoals begin/randvoorwaarden en coefficienten van de vergelijkingen (bijvoorbeeld de diffusiviteit in de warmtevergelijking). Het resultaat van de berekeningen is een lijst getallen corresponderend met de roosterpuntwaarden van de onbekende grootheden. Teneinde de resultaten beter te kunnen interpreteren worden zij meestal nog in een nabewerkings (post processing) fase grafies weergegeven. De numerieke berekeningen worden vanwege de grote omvang m.b.v. een computer uitgevoerd.

In dit college worden numerieke methoden behandeld die toepasbaar zijn voor het modelleren van uiteenlopende geodynamische processen. Onder numeriek modelleren verstaan we hier: het onderzoeken van modeleigenschappen, door de opgestelde modelvergelijkingen numeriek op te lossen voor verschillende waarden van de modelparameters. Technisch gezien komt dit neer op het herhaaldelijk oplossen van de betreffende partiële differentiaalvergelijkingen m.b.v. numeriek wiskundige discretisatie methoden.
Belangrijke geofysische modellerings problemen waarvoor numerieke methoden worden toegepast zijn:

- (Energie/massa) transportmodellen deze worden toegepast bij de bestudering van de warmtehuishouding in lithosfeermodellen of in stromingsmodellen toegepast op mantelconvectie of in modellen van poreuze media. Deze laatste modellen vinden op grote schaal toepassing in de hydrologie en aanverwante takken van millieu engineering en in de reservoir engineering i.v.m. de productie van olie en gas.

- Stromingsproblemen: - visceuze vloeistof modellen toegepast in modellen voor mantelconvectie en postglaciale opheffing.

- Deformatie problemen: elastische en plastische deformatie modellen worden toegepast in modellen op uiteenlopende schaal van gedetailleerde geologische modellen voor plooiende laagpakketten tot grootschalige lithosfeermodellen met meerdere lithosferische platen.

De hier te behandelen (geo)fysische toepassingen hebben betrekking op het numeriek oplossen van tijd- en plaatsafhankelijke vergelijkingen in de vorm van gekoppelde partiële differentiaalvergelijkingen. In het algemeen wordt het probleem eerst in de ruimte gediscretiseerd. Hierbij worden de ruimtelijke coordinaten als onafhankelijke variabelen vervangen door over te gaan op een eindig aantal discrete vrijheidsgraden. Hiermee vervangt men een partiële differentiaal vergelijking door een stelsel gewone differentiaalvergelijkingen met de tijd als enige onfhankelijke variabele. Deze stap wordt wel aangeduid als semi-discretisatie van het probleem. Het stelsel gewone differentiaalvergelijkingen wordt dan vervolgens in de tijd geïntegreerd m.b.v. een numerieke integratie methode. Bekende discretisatiemethoden zijn:

- eindige differentie- en eindige volumen methoden

- eindige elementen methoden

- spectraal methoden

De beide eerste categoriën zullen hier worden behandeld. Enkele integratiemethoden voor gewone differentiaalvergelijkingen zijn:

- Euler expliciet/impliciet ook wel aangeduid als Euler voorwaarts/achterwaarts

- Crank-Nicolson

- Runge-Kutta

De beide eerste methoden zullen hier worden toegepast.

28.2 Modelleren van 1-d warmte problemen

Als eerste voorbeeld van een probleem dat met numerieke methoden wordt gemodelleerd behandelen we het probleem van warmtediffusie in een statisch (niet stromend) medium. Dit probleem treedt o.a. op bij de bestudering van de warmtehuishouding van de lithosfeer van de aarde. In latere hoofdstukken zal aandacht worden besteed aan de behandeling van het complexer probleem van warmtetransport in een stromend medium dat optreedt bij de bestudering van thermische convectie in de aardmantel.

We beperken ons hier voor de overzichtelijkheid tot 1-D problemen. 1-D formulering van het warmteprobleem voor de lithosfeer is relevant omdat - op voldoende grote schaal - een horizontaal gelaagd model toepasbaar is waarin het warmtetransport voornamelijk in verticale richting plaatsvindt. De te behandelen oplossingsmethoden zijn echter zonder meer uit te breiden naar twee en drie dimensies. In hoofdstuk 3 worden vergelijkbare methoden behandeld voor 2-D problemen. We gaan uit van de 1-D tijdafhankelijke warmtevergelijking.

$$\frac{\partial T}{\partial t} \; = \; \nabla \cdot \kappa \nabla T \; + \; Q \; = \; \frac{\partial}{\partial z} \kappa \frac{\partial T}{\partial z} \; + \; Q \tag{28.1}$$

met T de temperatuur, κ de warmtediffusiviteit, $\kappa = k/\rho c_p$, k de thermische conductiviteit, ρ de massadichtheid en c_p de soortelijke warmte.

Een belangrijke toepassing van de 1-D vergelijking voor een statisch medium is het halfruimtemodel voor afkoelende en spreidende oceaanlithosfeer (Turcotte & Schubert, 1982). De advectieve term in de afgeleide naar de tijd in de warmtevergelijking voor het (horizontaal) stromend medium met snelheid u, $dT/dt = \partial T/\partial t + u\partial T/\partial x$ wordt hierin m.b.v. een ouderdomstranformatie $\tau = u/x$ omgezet in $\partial T/\partial \tau$. Gaan we daarnaast uit van een stationaire toestand met $\partial T/\partial t = 0$ en verwaarlozen we horizontale diffusie van warmte $\partial^2 T/\partial x^2 = 0$ dan resulteert de tijdafhankelijke vergelijking met de leeftijd τ in de rol van de tijdparameter.

Opgave:

Leid de bovenstaande warmte vergelijking af uit een behoudswet voor thermische energie voor een 1-D statisch medium.

Het probleem wordt opgesteld voor een 1-D oplossingsdomein, het interval $[z_0, z_{N+1}]$. Voor het tijdafhankelijke probleem veronderstellen we de temperatuur gegeven voor $t = 0$, d.m.v. de beginvoorwaarde

$$T(z, 0) = T_b(z) \tag{28.2}$$

Waarin $T_b(z)$ een bekende functie. We zullen twee typen randvoorwaarden behandelen (Dirichlet/Neumann) ook wel aangeduid als essentiële/natuurlijke randvoorwaarden,

$$T(z_r, t) = T_r(t) \ , \ type \ 1 \ (Dirichlet) \tag{28.3}$$

$$k\frac{\partial T(z_r, t)}{\partial z} = q_r(t) \ , \ type \ 2 \ (Neumann) \tag{28.4}$$

met q_r de voorgeschreven warmtestroom dichtheid. We behandelen twee verschillende discretisatie methoden. De eerste, gebaseerd op een centrale differentie benadering van de ruimtelijke afgeleiden in de warmtevergelijking, is beperkt toepasbaar. De tweede methode (eindige volume of box methode) is geschikt voor meer algemene problemen, bijvoorbeeld bij een variabele geleidingscoefficient k.

Voor het stationaire geval ($\partial T/\partial t = 0$) leiden beide discretisatie methoden tot een stelsel lineare algebraische vergelijkingen dat m.b.v. een computer numeriek kan worden opgelost. Voor het tijdafhankelijk geval ($\partial T/\partial t \neq 0$) resulteert een stelsel gewone differentiaalvergelijkingen met de tijd t als onafhankelijke variabele. In een later hoofdstuk zullen een aantal integratiemethoden worden behandeld voor dit stelsel gewone differentiaalvergelijkingen.

28.2.1 Een differentiemethode met equidistant rooster

We bespreken eerst een methode voor een equidistant rooster en met uniforme diffusie coefficient κ. Deze methode heeft als voordeel dat hij conceptueel eenvoudig is. In volgende secties wordt een algemener toepasbare methode behandeld.

Definieer een 1-D rooster met vaste afstand tussen de roosterpunten (equidistant):

$$z_i \ = \ z_0 + i \times \Delta z \ , \ i \ = \ 0, 1, 2, 3, ..., N + 1 \tag{28.5}$$

met randpunten z_0 en z_{N+1}, gedeeltelijk weergegeven in Fig.1.

Definieer vectoren $\mathbf{T}(t)$, $\mathbf{Q}(t)$ met als elementen de roosterpuntwaarden van de temperatuur $T_i(t) \ = \ T(z_i, t)$ en de warmteproductiviteit $Q_i(t) \ = \ Q(z_i, t)$. Voor het gemak laten we de aanduiding van de tijdafhankelijkheid weg in de volgende formules. Benader de afgeleide $\partial^2 T/\partial z^2$ in de warmtevergelijking m.b.v. een centrale differentie formule die kan worden afgeleid d.m.v. een Taylor reeks ontwikkeling van de temperatuur,

$$T(z + \Delta z) = T(z) + \Delta z\frac{\partial T}{\partial z} + \frac{\Delta z^2}{2}\frac{\partial^2 T}{\partial z^2} + \frac{\Delta z^3}{6}\frac{\partial^3 T}{\partial z^3} + \frac{\Delta z^4}{24}\frac{\partial^4 T}{\partial z^4} + ... \tag{28.6}$$

$$T(z \ - \ \Delta z) \ = \ T(z) \ - \ \Delta z \frac{\partial T}{\partial z} \ + \ \frac{\Delta z^2}{2} \frac{\partial^2 T}{\partial z^2} \ - \ \frac{\Delta z^3}{6} \frac{\partial^3 T}{\partial z^3} \ + \ \frac{\Delta z^4}{24} \frac{\partial^4 T}{\partial z^4} \ - \ ... \tag{28.7}$$

elimineer oneven machten van Δz:

$$T(z \ + \ \Delta z) \ - \ 2T(z) \ + \ T(z - \Delta z) \ = \ \Delta z^2 \frac{\partial^2 T}{\partial z^2} \ + \ \frac{\Delta z^4}{12} \frac{\partial^4 T}{\partial z^4} \ + \ ... \tag{28.8}$$

$$\frac{\partial^2 T}{\partial z^2} \ = \ \frac{T(z \ - \ \Delta z) \ - \ 2T(z) \ + \ T(z \ + \ \Delta z)}{\Delta z^2} \ - \ \frac{\Delta z^2}{12} \frac{\partial^4 T}{\partial z^4} \ + \ ... \tag{28.9}$$

Verwaarloozing in (9) van de termen in Δz^2, resulteert in een discrete warmtevergelijking voor een roosterpunt z_i:

$$\frac{\partial T_i}{\partial t} \ = \ \frac{\kappa}{\Delta z^2} [T_{i-1} \ - \ 2T_i \ + \ T_{i+1}] \ + \ Q_i \tag{28.10}$$

Bij gegeven temperatuur op de randen geldt: $T(z_0) = T_0$, $T(z_{N+1}) = T_{N+1}$. De vergelijkingen voor het eerste ($i = 1$) en laatste ($i = N$) inwendige roosterpunt krijgen dan een rechterlid bijdrage in de voorgeschreven randtemperatuur T_0 en T_{N+1}.

$$i \ = \ 1 \ \rightarrow \ \frac{\partial T_1}{\partial t} \ = \ \frac{\kappa}{\Delta z^2} [-2T_1 \ + \ T_2] \ + \ \frac{\kappa}{\Delta z^2} T_0 \ + \ Q_1 \tag{28.11}$$

$$i \ = \ N \ \rightarrow \ \frac{\partial T_N}{\partial t} \ = \ \frac{\kappa}{\Delta z^2} [T_{N-1} \ - \ 2T_N] \ + \ \frac{\kappa}{\Delta z^2} T_{N+1} \ + \ Q_N \tag{28.12}$$

E.e.a. resulteert in een stelsel van precies N gewone differentiaal vergelijkingen

$$\frac{\partial \mathbf{T}}{\partial t} \ = \ \mathbf{AT} \ + \ \mathbf{R} \tag{28.13}$$

met rechterlidvector:

$$\mathbf{R} \ = \ (Q_1 \ + \ \frac{\kappa}{\Delta z^2} T_0, Q_2, ..., Q_{N-1}, Q_N \ + \ \frac{\kappa}{\Delta z^2} T_{N+1}) \tag{28.14}$$

en $(N \times N)$ coefficienten matrix:

$$\mathbf{A} \ = \ -\frac{\kappa}{\Delta z^2} \begin{bmatrix} 2 & -1 & 0 & 0 & . & . & . & . \\ -1 & 2 & -1 & 0 & . & . & . & . \\ 0 & -1 & 2 & -1 & . & . & . & . \\ 0 & 0 & -1 & 2 & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & 2 & -1 & 0 \\ . & . & . & . & . & -1 & 2 & -1 \\ . & . & . & . & . & 0 & -1 & 2 \end{bmatrix} \tag{28.15}$$

en onbekende vector

$$\mathbf{T}(t) \ = \ (T_1(t), T_2(t), ..., T_N(t)) \tag{28.16}$$

Implementatie van natuurlijke randvoorwaarden

De behandeling van natuurlijke randvoorwaarden verschilt van die voor essentiele randvoorwaarden in de vorige sectie. We beschouwen een geval met voorgeschreven warmtestroomdichtheid q_{N+1} in het punt p_{N+1}. De temperatuur in p_{N+1} is nu ook een onbekende (vrijheidsgraad) van het gestelde probleem en we moeten de discrete vergelijking (10) ook in p_{N+1} evalueren om een volledig stel $(N+1) \times (N+1)$ vergelijkingen te krijgen.

Een eenvoudige methode om de voorgeschreven warmtestroom te implementeren is deze uit te drukken in de knooppuntwaarden d.m.v. een differentie benadering. We benaderen de temperatuur-gradient in de uitdrukking voor de warmtestroomdichtheid m.b.v. een centrale differentie

$$q_{N+1} = k \frac{\partial T}{\partial z}|_{z_{N+1}} = k \frac{T_{N+2} - T_N}{2\Delta z} + O(\Delta z^2) \tag{28.17}$$

waarin we gebruik maken van een virtueel roosterpunt p_{N+2} op afstand Δz buiten het oplossings-domein. Voor de temperatuur in het virtuele punt vinden we uit (17)

$$T_{N+2} = T_N + 2 \frac{q_{N+1}}{k} \Delta z \tag{28.18}$$

M.b.v. (18) kan de discrete vergelijking in het randpunt p_{N+1} worden geevalueerd resulterend in,

$$\frac{\partial T_{N+1}}{\partial t} = \frac{2\kappa}{\Delta z^2} [T_N - T_{N+1}] + \frac{2\kappa}{\Delta z} \frac{q_{N+1}}{k} + Q_{N+1} \tag{28.19}$$

We zien dat de inhomogene randvoorwaarde een extra bijdrage tot de rechterlidvector levert. Merk op dat dat de coefficienten matrix niet meer symmetrisch is. De symmetrie is echter eenvoudig te herstellen door de vergelijking voor het randpunt door 2 te delen.

Opgave:

Laat zien - door Taylor ontwikkeling - dat de gebruikte randvoorwaarde inderdaad van de tweede orde is in Δz. Ga na dat een meer voor de hand liggende voorwaartse differentie in een eerste orde nauwkeurigheid resulteert.

28.2.2   Een differentiemethode met variabele roosterafstand

Bij de afleiding van de eindige differentie methode voor de warmtevergelijking m.b.v. de centrale differentie benadering voor de tweede afgeleide term werd er vanuit gegaan dat het gebruikte rooster een vaste afstand tussen de roosterpunten heeft (equidistant rooster). We namen bovendien aan dat de warmtegeleidingscoefficient uniform was. Om voldoende nauwkeurigheid in de numerieke oplossing te bereiken kan het soms noodzakelijk zijn om veel roosterpunten te gebruiken wat direct resulteert in grotere rekentijd voor het computerprogramma en een groter beslag op het computergeheugen. Deze roosterverfijning wordt bij de equidistante aanpak over het gehele domein doorgevoerd terwijl deze in voorkomende gevallen uitsluitend op deelgebieden wenselijk is - bijvoorbeeld daar waar de oplossing sterk varieert.

We behandelen hier een methode met variabele roosterafstand waarmee het rooster locaal kan worden verfijnd. We zullen tegelijkertijd het algemene geval met variabele coefficient $k(z)$ bespreken. We beperken ons weer tot het 1-D geval. Generalisatie tot 2-D en 3-D toepassingen zijn direct mogelijk. De hier te behandelen methode is bekend als de 'eindige volume methode' of ook wel 'box methode'. We bespreken eerst het stationair probleem en zullen achteraf zien hoe het resultaat kan worden uitgebreid voor tijdafhankelijke problemen.

Discretisatie van de vergelijking

Bij de box methode wordt de partiële differentiaalvergelijking geintegreerd over het domein $I = [0, L]$, De integratie wordt uitgevoerd over deelintervallen I_i - de eindige volumen waaraan de methode zijn naam ontleent - gecentreerd rond roosterpunt p_i, $I_i = [z_{m_{i-1}}, z_{m_i}]$, zie Fig.2.

Hierin zijn de m_i de midpunten van de roosterintervallen. Integratie van de stationaire warmtevergelijking over I_i geeft

$$\int_{z_{m_{i-1}}}^{z_{m_i}} \left[\frac{d}{dz} k(z) \frac{d}{dz} T \; + \; H(z) \right] dz \; =$$
$$k(z) \frac{dT}{dz} |_{z_{m_{i-1}}}^{z_{m_i}} \; + \; \int_{z_{m_{i-1}}}^{z_{m_i}} H(z) dz = 0 \tag{28.20}$$

Merk op dat de overgebleven afgeleide in (1) wordt geevalueerd in de midpunten m_{i-1} en m_i. We benaderen vervolgens deze afgeleide in de midpunten door een centrale differentie in termen van de aangrenzende knooppunt waarden en we definieren nog $k(z_{m_i}) \; = \; k_i$ en $z_{p_{i+1}} \; - \; z_{p_i} \; = \; h_i$.

$$k \frac{dT}{dz} |_{z_{m_i}} \approx k_i \frac{T_{i+1} \; - \; T_i}{z_{p_{i+1}} \; - \; z_{p_i}} \; = \; k_i \frac{T_{i+1} \; - \; T_i}{h_i} \tag{28.21}$$

$$k \frac{dT}{dz} |_{z_{m_{i-1}}} \approx k_{i-1} \frac{T_i \; - \; T_{i-1}}{z_{p_i} \; - \; z_{p_{i-1}}} \; = \; k_{i-1} \frac{T_i \; - \; T_{i-1}}{h_{i-1}} \tag{28.22}$$

De warmteproduktie wordt bekend verondersteld en we definieren

$$\int_{z_{m_{i-1}}}^{z_{m_i}} H(z) dz \; = \; F_i \tag{28.23}$$

Invullen van (2),(3) en (4) in (1) levert,

$$-\frac{k_{i-1}}{h_{i-1}} T_{i-1} \; + \; \left[\frac{k_{i-1}}{h_{i-1}} \; + \; \frac{k_i}{h_i} \right] T_i \; - \; \frac{k_i}{h_i} T_{i+1} = F_i \tag{28.24}$$

Vergelijking (5) is een lineaire algebraische vergelijking in de onbekende roosterpuntenwaarden van de temperatuur T_i. Door het beschreven integratie procede te herhalen voor alle - N - eindige volumen I_i krijgen we een stelsel lineaire vergelijkingen.

Opgave:
Ga na dat het stelsel volledig is als we de temperatuur op de randen van het interval $[0, L]$ voorschrijven m.b.v. twee essentiele randvoorwaarden. Volledig betekent hier dat er evenveel vergelijkingen als onbekenden zijn.

Opgave:
Laat zien dat het stelsel (5) voor het speciaal geval met een equidistant rooster en uniforme coefficient k identiek is aan het in de voorgaande sectie afgeleide stelsel 2.1(13). Hoe verhouden de rechterlidvectoren zich ?

Het stelsel vergelijkingen in matrix vorm geschreven is

$$\mathbf{A T} = \mathbf{F} \tag{28.25}$$

Opgave:
Toon aan dat de matrix $\mathbf{A}$ symmetrisch en tri-diagonaal is, d.w.z. $A_{ij} \; = \; A_{ji}$, $A_{ij} \; = \; 0, |i \; - \; j| > 1$.

Opgave:

Ga na dat de bovenstaande formulering ook nog toepasbaar is als we de coefficient van de differentiaalvergelijking $k(z)$ definieren d.m.v. een rij i.h.a. verschillende rooster segmentwaarden d.w.z. als we definieren,

$$k(z) = k_i, \; z_{p_i} < z < z_{p_{i+1}} \tag{28.26}$$

Hiermee kunnen gelaagde modellen met discontinuiteiten in de conductiviteit worden gemodelleerd.

Opgave:

Als we de warmteproduktie coefficient $H(z)$ op dezelfde wijze definieren als in de geleidingscoefficient $k(z)$ in bovenstaande opgave (stukwijs constant) dan kan voor de rechterlidvector $\mathbf{F}$ in (4) worden afgeleid (ga na)

$$F_i = \int_{z_{m_{i-1}}}^{z_{m_i}} H(z)dz = \frac{h_{i-1}}{2}H(z_{m_{i-1}}) + \frac{h_i}{2}H(z_{m_i}) \tag{28.27}$$

Opgave:

Ga na: wanneer de warmteproduktie geconcentreerd is in een vlak $z = z_s$, $H(z) = H_s\delta(z - z_s)$ met $z_{m_{k-1}} < z_s < z_{m_k}$ - d.w.z. $z_s \in I_k$, dan geldt voor de rechterlidvector in (4) $F_i = H_s\delta_{ik}$ met δ_{ik} het Kronecker delta symbool.

Behandeling van randvoorwaarden

De behandeling van essentiele randvoorwaarden gaat op een vergelijkbare manier als bij de discretisatie methode van sectie 2.1. Essentiele randvoorwaarden geven een bijdrage tot de rechterlidvector en ze reduceren het aantal vrijheidsgraden. We behandelen hier de implementatie van een natuurlijke of Neumann randvoorwaarde die optreedt als de warmtestroomdichtheid op de rand is voorgeschreven. We stellen $kdT/dz|_{z=z_{N+1}} = q_{N+1}$. De behandeling verschilt van de in sectie 2.1.1 voor een equidistant rooster beschreven methode. We werken de behandeling uit van de randvoorwaarde in roosterpunt p_{N+1} - zie Fig. 2.3.

In dit geval is de temperatuur in het randpunt p_{N+1} ook een vrijheidsgraad. Teneinde een volledig stel vergelijkingen te krijgen moet de discretisatie 2.2(1) ook in het roosterpunt p_{N+1} geevalueerd worden. We werken de evaluatie uit voor de roosterpunten p_N en p_{N+1}.

Voor p_N krijgen we m.b.v. 2.2.1(5)

$$-\frac{k_{N-1}}{h_{N-1}}T_{N-1} + \left[\frac{k_{N-1}}{h_{N-1}} + \frac{k_N}{h_N}\right]T_N - \frac{k_N}{h_N}T_{N+1} = F_N \tag{28.28}$$

De randtemperatuur T_{N+1} treedt nu als vrijheidsgraad van het probleem in het linkerlid op.

Voor het randpunt met voorgeschreven warmtestroomdichtheid p_{N+1} gaan we uit van integraaluitdrukking 2.2(1) en we kiezen het integratieinterval $I_{N+1} = [z_{m_N}, z_{p_{N+1}}]$,

$$\int_{z_{m_N}}^{z_{p_{N+1}}}\left[\frac{d}{dz}k\frac{d}{dz}T + H\right]dz =$$

$$k_N\frac{dT}{dz}|_{z_{p_{N+1}}} - k_N\frac{dT}{dz}|_{z_{m_N}} + \int_{z_{m_N}}^{z_{p_{N+1}}}H(z)dz \approx$$

$$q_{N+1} - k_N\frac{T_{N+1} - T_N}{h_N} + F_{N+1} = 0 \tag{28.29}$$

Waarmee we voor de vergelijking voor roosterpunt p_{N+1} vinden

$$-\frac{k_N}{h_N}T_N + \frac{k_N}{h_N}T_{N+1} = F_{N+1} + q_{N+1} \tag{28.30}$$

We zien dat de voorgeschreven warmtestroomdichtheid q_{N+1} als extra bijdrage in de rechterlid vector verschijnt.

Appendix A

Don't be a hero - unless you have to

What follows was published online on July 17th, 2017 at `https://blogs.egu.eu/divisions/gd/2017/07/19/dont-be-a-hero-unless-you-have-to/` It was written by me and edited by Iris van Zelst, at the time PhD student at ETH Zürich.

In December 2013, I was invited to give a talk about the Aspect code [1] at the American Geological Union conference in San Francisco. Right after my talk, Prof. Louis Moresi took the stage and gave a talk entitled: *Underworld: What we set out to do, How far did we get, What did we Learn?*

The abstract went as follows:

"Underworld was conceived as a tool for modelling 3D lithospheric deformation coupled with the underlying / surrounding mantle flow. The challenges involved were to find a method capable of representing the complicated, non-linear, history dependent rheology of the near surface as well as being able to model mantle convection, and, simultaneously, to be able to solve the numerical system efficiently. [...] The elegance of the method is that it can be completely described in a couple of sentences. However, there are some limitations: it is not obvious how to retain this elegance for unstructured or adaptive meshes, arbitrary element types are not sufficiently well integrated by the simple quadrature approach, and swarms of particles representing volumes are usually an inefficient representation of surfaces."

Aside from the standard numerical modelling jargon, Louis used a term during his talk which I thought at the time had a nice ring to it: hero codes. In short, I believe he meant the codes written essentially by one or two people who at some point in time spent great effort into writing a code (usually choosing a range of applications, a geometry, a number of dimensions, a particular numerical method to solve the relevant PDEs(1), and a tracking method for the various fields of interest).

In the long list of Hero codes, one could cite (in alphabetical order) CitcomS [1], Douar [8], Fantom [2], IELVIS [5], LaMEM [3], pTatin [4], SLIM3D [10], Sopale [7], StaggYY [6], Sulec [11], Underworld [9], and I apologise to all other heroes out there whom I may have overlooked. And who does not want to be a hero? The Spiderman of geodynamics, the Superwoman of modelling?

Louis' talk echoed my thoughts on two key choices we (computational geodynamicists) are facing: Hero or not, and if yes, what type?

Hero or not?

Speaking from experience, it is an intense source of satisfaction when peer-reviewed published results are obtained with the very code one has painstakingly put together over months, if not years. But is it worth it?

On the one hand, writing one own's code is a source of deep learning, a way to ensure that one understands the tool and knows its limitations, and a way to ensure that the code has the appropriate combination of features which are necessary to answer the research question at hand. On the other hand, it is akin to a journey; a rather long term commitment; a sometimes frustrating endeavour,

1249

with no guarantee of success. Let us not deny it - many a student has started with one code only to switch to plan B sooner or later. Ultimately, this yields a satisfactory tool with often little to no perennial survival over the 5 year mark, a scarce if at all existent documentation, and almost always not compliant with the growing trend of long term repeatability. Furthermore, the resulting code will probably bear the marks of its not-all-knowing creator in its DNA and is likely not to be optimal nor efficient by modern computational standards.

This brings me to the second choice: elegance & modularity or taylored code & raw performance? Should one develop a code in a very broad framework using as much external libraries as possible or is there still space for true heroism?

It is my opinion that the answer to this question is: both. The current form of heroism no more lies in writing one's own FEM(2)/FDM(3) packages, meshers, or solvers from scratch, but in cleverly taking advantage of state-of-the-art packages such as for example p4est [15] for Adaptive Mesh Refinement, PetSc [13] or Trilinos [14] for solvers, Saint Germain [17] for particle tracking, deal.ii [12] or Fenics [16] for FEM, and sharing their codes through platforms such as svn, bitbucket or github.

In reality, the many different ways of approaching the development or usage of a (new) code is linked to the diversity of individual projects, but ultimately anyone who dares to touch a code (let alone write one) is a hero in his/her own right: although (super-)heroes can be awesome on their own, they often complete each other, team up and join forces for maximum efficiency. Let us all be heroes, then, and join efforts to serve Science to the best of our abilities.

Abbreviations
    (1) PDE: Partial Differential Equation
    (2) FEM: Finite Element Method
    (3) FDM: Finite Difference Method

References

[1] Zhong *et al.* , JGR 105, 2000;

[2] Thieulot, PEPI 188, 2011;

[3] Kaus *et al.* , NIC Symposium proceedings, 2016;

[4] May *et al.* , CMAME 290, 2015

[5] Gerya and Yuen, PEPI 163, 2007

[6] Tackley, PEPI 171, 2008

[7] Fullsack, GJI 120, 1995

[8] Braun *et al.* , PEPI 171, 2008

[9] http://www.underworldcode.org/

[10] Popov and Sobolev, PEPI 171, 2008

[11] http://www.geodynamics.no/buiter/sulec.html

[12] Bangerth *et al.* , J. Numer. Math., 2016; http://www.dealii.org/

[13] http://www.mcs.anl.gov/petsc/

[14] https://trilinos.org/

[15] Burstedde *et al.* , SIAM journal on Scientific Computing, 2011; http://www.p4est.org/

[16] https://fenicsproject.org/

[17] Quenette *et al.* , Proceedings 19th IEEE, 2007

Appendix B

Some useful Python commands

B.0.1 Sparse matrices

So far, the best way I have found to deal with sparse matrices is to declare the matrix as a 'lil_matrix' (linked list).

```python
from scipy.sparse import csr_matrix, lil_matrix
A_mat = lil_matrix((Nfem,Nfem),dtype=np.float64)
```

One then adds terms to it as if it was a full/dense matrix. Once the assembly is done, the conversion to CSR format is trivial:

```python
A_mat=A_mat.tocsr()
```

Finally the solver can be called:

```python
sol=sps.linalg.spsolve(A_mat,rhs)
```

B.0.2 condition number

if the matrix has been declared as lil_matrix, first convert it to a dense matrix:

```python
A_mat=A_mat.dense()
```

The condition number of the matrix is simply obtained as follows:

```python
from numpy import linalg as LA
print(LA.cond(A_mat))
```

B.0.3 Weird stuff

Python is touted as the one language students should learn and master. However it is a language which allows *way* too much liberty in its syntax and encourages students to be sloppy.

For instance the following code runs just fine:

```python
for k in range(0,5):
    for k in range(0,5):
        for k in range(0,5):
            print(k)
```

This alone should disqualify this language. It is easy to see the obvious problem with this code, but adding a few lines of code in between each 'for' line hides the problem and the absence of any warning makes this code a nightmare to debug.

Here is another example:

```python
import numpy as np

Told=np.full(10,123)

Tnew=np.zeros(10)
for i in range(0,10):
    Tnew[i]=np.sqrt(i**5)

print(Tnew)

Told[:]=Tnew[:]

print(Told)
```

If Told is not defined explicitly as a float, then the numer 123 will decide of its type (in this case integer). If '123' is changed into '123.' then the array will be initialised as an array of floats. In the first case

```
Told is [  0   1   5  15  32  55  88 129 181 243]
```

while in the second case it is equal to

```
[  0.          1.          5.65685425  15.58845727  32.
  55.90169944  88.18163074 129.64181424 181.01933598 243.        ]
```

This is very dangerous since our students are not taught number representation.

https://www.w3schools.com/python/default.asp

https://www.codecademy.com/learn/learn-python

https://learnpythonthehardway.org/book/

B.0.4   Making simple 2D plots

```python
import matplotlib.pyplot as plt

# number of points
N=100

# a despicable way of filling two arrays
x_data=[]
y_data=[]
for i in range(0,N):
    x=i
    y=i**2+2*i+1.
    x_data.append(x)
    y_data.append(y)

# generating a 2D figure with the data
plt.figure()
plt.plot(x_data,y_data, label = 'name_of_data')
plt.xlabel('x-axis_label')
```

```
plt.ylabel('y-axis label')
plt.legend()
plt.savefig('myplot.pdf', bbox_inches='tight')
plt.show()
```
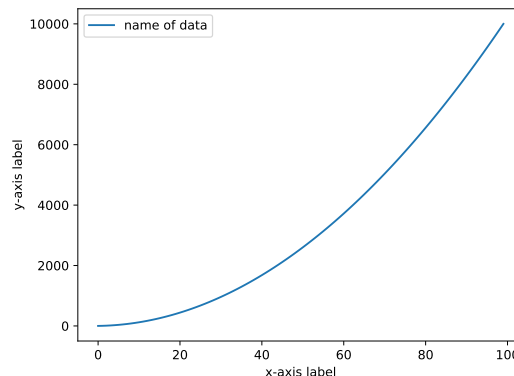

B.0.5 Making simple 3D plots of scatter

```
fig = plt.figure()
ax = plt.axes(projection='3d')
ax.set_title("insert here text for title")
size = ..some value..
ax.scatter3D(x, y, z, s = size)
```

B.0.6 How to debug your code

Debugging a FE code is by no means trivial. There is (at least) a grid, a connectivity array, basis functions and their derivatives, the elemental matrices and rhs, the assembly, boundary conditions, and a call to a solver before the solution (if the solver returns one!) can be visualised.

- First and foremost, make sure that your grid of points is correct. For instance, you can resort to exporting it to an ascii file as follows:

```
np.savetxt('velocity.ascii',np.array([x,y,u,v]).T,header='# x,y,u,v')
```

  In two dimensions, you should set for example nelx=3 and nely=2, so that for Q_1 elements the grid counts 12 points. Then make sure the coordinates and the order of the points makes sense. Repeat the process for pressure nodes, temperature nodes, etc ...

- Then it is time to check the connectivity array(s).

```
for iel in range (0,nel):
    print ("iel=",iel)
    for k in range(0,m):
        print ("node ",icon[0,iel],"at pos.",x[icon[0,iel]], y[icon[0,iel]])
```

  This displays the list of nodes and their positions making each element. Repeat the process for every connectivity array.

- We can go on with testing that the all basis functions are 1 on their node and zero elsewhere:

```
for i in range(0,m):
    print ('node',i,':',NNV(rnodes[i],snodes[i]))
```

here the arrays rnodes and snodes contain the (r,s) coordinates of the m nodes

- test jacobian, compute volume of domain

- sum(dNNNdx)=0

- print nodes where bc

B.0.7   Optional arguments

Courtesy of Henry Brett.

```python
def myfunc(a,b, *optional_arguments, **keyword_arguments):
    print(a)
    print(b)
    for ar in optional_arguments:
        print(ar)
    d=keyword_arguments.get("d", None)
    print(d)


a="dog"
b="cat"
myfunc(a,b,"shrek","fiona",d="donkey")
```

B.0.8   drawing and filling quadrilaterals

```python
import matplotlib.pyplot as plt

plt.figure(figsize=(3, 3))
plt.axis('equal')

x=(1,1.5,1.6,0.8)
y=(0,-0.1,1.5,1.2)
plt.fill(x, y,"b")

x=(0,0.5,0.6,-0.2)
y=(0,-0.4,1.1,1.1)
plt.fill(x, y,"r")

plt.savefig('xxx.pdf')
plt.show()
```
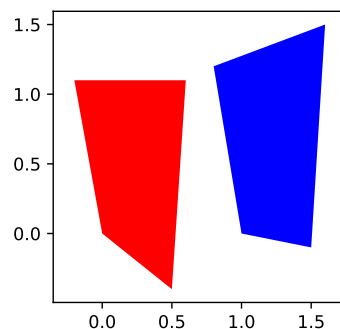
Appendix C

Elemental matrices for simple geometries

In what follows I compute the mass matrix for a variety of reference elements. If you wish to use these in a code, do not forget to take the jacobian of the transformation/mapping into account.

C.0.1 1D segments

Linear basis functions

Let us start with the mass matrix (which we encountered in Section 9.1 – although we leave the ρC_p term out):

$$\boldsymbol{M}_e = \int_{\Omega_e} \vec{N}^T \vec{N} dV = \int_{-1}^{+1} \vec{N}^T \vec{N} dr \tag{C.1}$$

on the reference element, with

$$\vec{N}^T = \left(\begin{array}{c} N_1(r) \\ N_2(r) \end{array} \right) = \frac{1}{2} \left(\begin{array}{c} 1 - r \\ 1 + r \end{array} \right)$$

We have

$$\int_{-1}^{+1} N_1(r)N_1(r)dr = 2/3 \tag{C.2}$$

$$\int_{-1}^{+1} N_1(r)N_2(r)dr = 1/3 \tag{C.3}$$

$$\int_{-1}^{+1} N_2(r)N_2(r)dr = 2/3 \tag{C.4}$$

Following the procedure in Section 9.1 we arrive at

$$\boldsymbol{M}^e = \frac{1}{3} \left(\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array} \right)$$

The lumped mass matrix is then

$$\bar{\boldsymbol{M}}^e = \frac{1}{3} \left(\begin{array}{cc} 2 + 1 & 0 \\ 0 & 1 + 2 \end{array} \right) = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right) \tag{C.5}$$

Remark. *The sum of all the terms in the mass matrix must be equal to 2. Indeed:*

$$
\begin{aligned}
\sum_{ij} M_{ij} &= \sum_{ij} \int_{-1}^{+1} N_i N_j dr \\
&= \int_{-1}^{+1} (N_1 N_1 + N_1 N_2 + N_2 N_1 + N_2 N_2) dr \\
&= \int_{-1}^{+1} [N_1(N_1 + N_2) + N_2(N_1 + N_2)] dr \\
&= \int_{-1}^{+1} (N_1 + N_2) dr \\
&= 2
\end{aligned}
$$

Quadratic basis functions

There are now three nodes in the segment so that the mass matrix is now a 3×3 matrix. We have (see Section 8.2.1)

$$
\vec{N}^T(r) = \begin{pmatrix} N_1(r) \\ N_2(r) \\ N_3(r) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}r(r-1) \\ 1 - r^2 \\ \frac{1}{2}r(r+1) \end{pmatrix}
\tag{C.6}
$$

We then have to compute

$$
\begin{aligned}
\int_{-1}^{+1} N_1(r) N_1(r) dr &= \frac{8}{30} = 0.26666 \\
\int_{-1}^{+1} N_1(r) N_2(r) dr &= \frac{4}{30} = 0.13333 \\
\int_{-1}^{+1} N_1(r) N_3(r) dr &= -\frac{2}{30} = -0.06666... \\
\int_{-1}^{+1} N_2(r) N_2(r) dr &= \frac{16}{15} = 1.06666 \\
\int_{-1}^{+1} N_2(r) N_3(r) dr &= \frac{4}{30} = 0.13333 \\
\int_{-1}^{+1} N_3(r) N_3(r) dr &= \frac{8}{30} = 0.26666
\end{aligned}
$$

and finally

$$
\boldsymbol{M}^e = \frac{1}{30} \begin{pmatrix} 8 & 4 & -2 \\ 4 & 32 & 4 \\ -2 & 4 & 8 \end{pmatrix}
\tag{C.7}
$$

The lumped mass matrix is then

$$
\begin{aligned}
\bar{\boldsymbol{M}}^e &= \frac{1}{30} \begin{pmatrix} 8 + 4 - 2 & 0 & 0 \\ 0 & 4 + 32 + 4 & 0 \\ 0 & 0 & -2 + 4 + 8 \end{pmatrix} \\
&= \frac{1}{30} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 40 & 0 \\ 0 & 0 & 10 \end{pmatrix} \\
&= \frac{1}{3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}
\end{aligned}
\tag{C.8}
$$

We can easily verify that

$$\sum_{ij} M_{ij} = 2 \qquad\qquad \sum_{ij} \bar{M}_{ij} = 2$$

Cubic basis functions

There are now four nodes in the segment so that the mass matrix is now a 4×4 matrix. We have (see Section 8.2.3)

$$\vec{N}^T(r) = \begin{pmatrix} N_1(r) \\ N_2(r) \\ N_3(r) \\ N_4(r) \end{pmatrix} = \frac{1}{16} \begin{pmatrix} -1 + r + 9r^2 - 9r^3 \\ 9 - 27r - 9r^2 + 27r^3 \\ 9 + 27r - 9r^2 - 27r^3 \\ -1 - r + 9r^2 + 9r^3 \end{pmatrix} \tag{C.9}$$

$$\int_{-1}^{+1} N_1(r) N_1(r) dr = \frac{1}{256} \frac{4096}{105}$$

$$\int_{-1}^{+1} N_1(r) N_2(r) dr = \frac{1}{256} \frac{1056}{35}$$

$$\int_{-1}^{+1} N_1(r) N_3(r) dr = -\frac{1}{256} \frac{384}{35}$$

$$\int_{-1}^{+1} N_1(r) N_4(r) dr = \frac{1}{256} \frac{608}{105}$$

$$\int_{-1}^{+1} N_2(r) N_2(r) dr = \frac{1}{256} \frac{6912}{35}$$

$$\int_{-1}^{+1} N_2(r) N_3(r) dr = -\frac{1}{256} \frac{864}{35}$$

$$\int_{-1}^{+1} N_2(r) N_4(r) dr = -\frac{1}{256} \frac{384}{35}$$

$$\int_{-1}^{+1} N_3(r) N_3(r) dr = \frac{1}{256} \frac{6912}{35}$$

$$\int_{-1}^{+1} N_3(r) N_4(r) dr = \frac{1}{256} \frac{1056}{35}$$

$$\int_{-1}^{+1} N_4(r) N_4(r) dr = \frac{1}{256} \frac{4096}{105}$$

and finally

$$\boldsymbol{M}^e = \frac{1}{16} \frac{1}{105} \begin{pmatrix} 256 & 198 & -72 & 38 \\ 198 & 1296 & -162 & -72 \\ -72 & -162 & 1296 & 198 \\ 38 & -72 & 198 & 256 \end{pmatrix} \tag{C.10}$$

1257

The lumped mass matrix is then

$$
\bar{M}^e = \frac{1}{16}\frac{1}{105}
\begin{pmatrix}
256 + 198 - 72 + 38 & 0 & 0 & 0 \\
0 & 198 + 1296 - 162 - 72 & 0 & 0 \\
0 & 0 & -72 - 162 + 1296 + 198 & 0 \\
0 & 0 & 0 & 38 - 72 + 198 +
\end{pmatrix}
$$

$$
= \frac{1}{16}\frac{1}{105}
\begin{pmatrix}
420 & 0 & 0 & 0 \\
0 & 1260 & 0 & 0 \\
0 & 0 & 1260 & 0 \\
0 & 0 & 0 & 420
\end{pmatrix}
$$

$$
= \frac{1}{4}
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 3 & 0 & 0 \\
0 & 0 & 3 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
$$

We can easily verify that

$$
\sum_{ij} M_{ij} = 2 \qquad \sum_{ij} \bar{M}_{ij} = 2
$$

Quartic basis functions

There are now five nodes in the segment so that the mass matrix is now a 5×5 matrix. We have (see Section 8.2.4)

$$
\vec{N}^T(r) =
\begin{pmatrix}
N_1(r) \\
N_2(r) \\
N_3(r) \\
N_4(r) \\
N_5(r)
\end{pmatrix}
= \frac{1}{6}
\begin{pmatrix}
r - r^2 - 4r^3 + 4r^4 \\
-8r + 16r^2 + 8r^3 - 16r^4 \\
6 - 30r^2 + 24r^4 \\
8r + 16r^2 - 8r^3 - 16r^4 \\
-r - r^2 + 4r^3 + 4r^4
\end{pmatrix}
\tag{C.11}
$$

$$\int_{-1}^{+1} N_1(r)N_1(r)dr = \frac{1}{36}\frac{1168}{315}$$

$$\int_{-1}^{+1} N_1(r)N_2(r)dr = \frac{1}{36}\frac{1184}{315}$$

$$\int_{-1}^{+1} N_1(r)N_3(r)dr = -\frac{1}{36}\frac{232}{105}$$

$$\int_{-1}^{+1} N_1(r)N_4(r)dr = \frac{1}{36}\frac{32}{45}$$

$$\int_{-1}^{+1} N_1(r)N_5(r)dr = -\frac{1}{36}\frac{116}{315}$$

$$\int_{-1}^{+1} N_2(r)N_2(r)dr = \frac{1}{36}\frac{1024}{45}$$

$$\int_{-1}^{+1} N_2(r)N_3(r)dr = -\frac{1}{36}\frac{512}{105}$$

$$\int_{-1}^{+1} N_2(r)N_4(r)dr = \frac{1}{36}\frac{1024}{315}$$

$$\int_{-1}^{+1} N_2(r)N_5(r)dr = \frac{1}{36}\frac{32}{45}$$

$$\int_{-1}^{+1} N_3(r)N_3(r)dr = \frac{1}{36}\frac{832}{35}$$

$$\int_{-1}^{+1} N_3(r)N_4(r)dr = -\frac{1}{36}\frac{512}{105}$$

$$\int_{-1}^{+1} N_3(r)N_5(r)dr = -\frac{1}{36}\frac{232}{105}$$

$$\int_{-1}^{+1} N_4(r)N_4(r)dr = \frac{1}{36}\frac{1024}{45}$$

$$\int_{-1}^{+1} N_4(r)N_5(r)dr = \frac{1}{36}\frac{1184}{315}$$

$$\int_{-1}^{+1} N_5(r)N_5(r)dr = \frac{1}{36}\frac{1168}{315} \tag{C.12}$$

$$\boldsymbol{M}^e = \frac{1}{36}\frac{1}{315}\begin{pmatrix} 1168 & 1184 & -696 & 224 & -116 \\ 1184 & 7168 & -1536 & 1024 & 224 \\ -696 & -1536 & 7488 & -1536 & -696 \\ 224 & 1024 & -1536 & 7168 & 1184 \\ -116 & 224 & -696 & 1184 & 1168 \end{pmatrix} \tag{C.13}$$

The lumped mass matrix is then

$$\bar{\boldsymbol{M}}^e = = \frac{1}{36}\frac{1}{315}\begin{pmatrix} 1764 & 0 & 0 & 0 & 0 \\ 0 & 8064 & 0 & 0 & 0 \\ 0 & 0 & 3024 & 0 & 0 \\ 0 & 0 & 0 & 8064 & 0 \\ 0 & 0 & 0 & 0 & 1764 \end{pmatrix} = \frac{1}{45}\begin{pmatrix} 7 & 0 & 0 & 0 & 0 \\ 0 & 32 & 0 & 0 & 0 \\ 0 & 0 & 12 & 0 & 0 \\ 0 & 0 & 0 & 32 & 0 \\ 0 & 0 & 0 & 0 & 7 \end{pmatrix} \tag{C.14}$$

We can once again easily verify that

$$\sum_{ij} M_{ij} = 2 \qquad \sum_{ij} \bar{M}_{ij} = 2$$

Note that all the integrals above were done very conveniently with the WolframAlpha software/website[1]. Example:
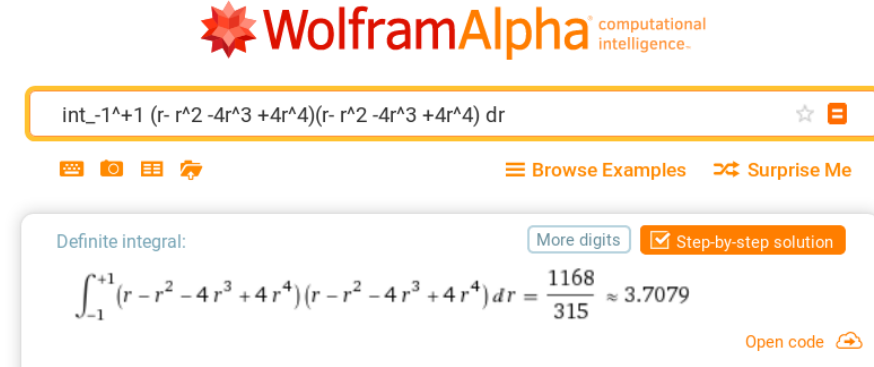
C.0.2   Quadrilaterals: rectangular linear elements

Mass matrix

We assume that each element is a rectangle of size $h_x \times h_y$. We start from the linear basis functions in the reference element as a function of r, s:

$$N_1(r, s) = \frac{1}{4}(1 - r)(1 - s) \tag{C.15}$$

$$N_2(r, s) = \frac{1}{4}(1 + r)(1 - s) \tag{C.16}$$

$$N_3(r, s) = \frac{1}{4}(1 + r)(1 + s) \tag{C.17}$$

$$N_3(r, s) = \frac{1}{4}(1 - r)(1 + s) \tag{C.18}$$

and their derivatives:

$$\partial_r N_1(r, s) = -\frac{1}{4}(1 - s)$$

$$\partial_r N_2(r, s) = \frac{1}{4}(1 - s)$$

$$\partial_r N_3(r, s) = \frac{1}{4}(1 + s)$$

$$\partial_r N_4(r, s) = -\frac{1}{4}(1 + s)$$

$$\partial_s N_1(r, s) = -\frac{1}{4}(1 - r)$$

$$\partial_s N_2(r, s) = -\frac{1}{4}(1 + r)$$

$$\partial_s N_3(r, s) = \frac{1}{4}(1 + r)$$

$$\partial_s N_4(r, s) = \frac{1}{4}(1 - r)$$

---

[1] https://www.wolframalpha.com/

We wish to compute the integral of a function $f(x, y)$ over the rectangular element:

$$\iint f(x, y)\,dx\,dy = \iint f(x(r, s), y(r, s)) \left| \frac{\partial(x, y)}{\partial(r, s)} \right| dr\,ds \tag{C.19}$$

$$= \iint f(x(r, s), y(r, s)) \begin{vmatrix} \partial x/\partial r & \partial x/\partial s \\ \partial y/\partial r & \partial y/\partial s \end{vmatrix} dr\,ds \tag{C.20}$$

From

$$x(r, s) = \sum_{i=1}^{4} N_i(r, s)x_i \qquad \text{andd} \qquad y(r, s) = \sum_{i=1}^{4} N_i(r, s)y_i$$

we can write

$$\frac{\partial x}{\partial r}(r, s) = \frac{\partial N_1}{\partial r}x_1 + \frac{\partial N_2}{\partial r}x_2 + \frac{\partial N_3}{\partial r}x_3 + \frac{\partial N_4}{\partial r}x_4$$

$$= -\frac{1}{4}(1 - s)x_1 + \frac{1}{4}(1 - s)x_2 + \frac{1}{4}(1 + s)x_3 - \frac{1}{4}(1 + s)x_4$$

$$= \frac{1}{4}(-x_1 + x_2 + x_3 - x_4 + s(x_1 - x_2 + x_3 - x_4))$$

$$= \frac{1}{4}(h_x + h_x + s(x_1 - x_2 + x_2 - x_1))$$

$$= \frac{1}{2}h_x$$

$$\frac{\partial x}{\partial s}(r, s) = \frac{\partial N_1}{\partial s}x_1 + \frac{\partial N_2}{\partial s}x_2 + \frac{\partial N_3}{\partial s}x_3 + \frac{\partial N_4}{\partial s}x_4$$

$$= -\frac{1}{4}(1 - r)x_1 - \frac{1}{4}(1 + r)x_2 + \frac{1}{4}(1 + r)x_3 + \frac{1}{4}(1 - r)x_4$$

$$= \frac{1}{4}(-x_1 - x_2 + x_3 + x_4 + r(x_1 - x_2 + x_3 - x_4))$$

$$= \frac{1}{4}(-x_1 - x_2 + x_2 + x_1 + r(x_1 - x_2 + x_2 - x_1))$$

$$= 0$$

$$\frac{\partial y}{\partial r}(r, s) = 0$$

$$\frac{\partial y}{\partial s}(r, s) = \frac{1}{2}h_y \tag{C.21}$$

Then

$$\begin{vmatrix} \partial x/\partial r & \partial x/\partial s \\ \partial y/\partial r & \partial y/\partial s \end{vmatrix} = \begin{vmatrix} h_x & 0 \\ 0 & h_y \end{vmatrix} = \frac{h_x h_y}{4}$$

and finally

$$\boxed{\iint_{\square} f(x, y)\,dx\,dy = \frac{h_x h_y}{4} \int_{-1}^{1} \int_{-1}^{1} f(x(r, s), y(r, s))\,dr\,ds} \tag{C.22}$$

Then the mass matrix is given by

$$
\boldsymbol{M}_e = \frac{h_x h_y}{4} \int_{-1}^{1} \int_{-1}^{1} \begin{pmatrix} N_1(r,s)N_1(r,s) & N_1(r,s)N_2(r,s) & N_1(r,s)N_3(r,s) & N_1(r,s)N_4(r,s) \\ N_2(r,s)N_1(r,s) & N_2(r,s)N_2(r,s) & N_2(r,s)N_3(r,s) & N_2(r,s)N_4(r,s) \\ N_3(r,s)N_1(r,s) & N_3(r,s)N_2(r,s) & N_3(r,s)N_3(r,s) & N_3(r,s)N_4(r,s) \\ N_4(r,s)N_1(r,s) & N_4(r,s)N_2(r,s) & N_4(r,s)N_3(r,s) & N_4(r,s)N_4(r,s) \end{pmatrix} drds
$$

$$
= \frac{h_x h_y}{9} \begin{pmatrix} 1 & 1/2 & 1/4 & 1/2 \\ 1/2 & 1 & 1/2 & 1/4 \\ 1/4 & 1/2 & 1 & 1/2 \\ 1/2 & 1/4 & 1/2 & 1 \end{pmatrix} \tag{C.23}
$$

Diffusion matrix

$$
\boldsymbol{K}_d^e = k \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \boldsymbol{B}^T(r,s) \cdot \boldsymbol{B}(r,s)\, drds
$$

with

$$
\boldsymbol{B}(r,s) = \begin{pmatrix} -\frac{1}{h_x}\frac{1}{2}(1-s) & \frac{1}{h_x}\frac{1}{2}(1-s) & \frac{1}{h_x}\frac{1}{2}(1+s) & -\frac{1}{h_x}\frac{1}{2}(1+s) \\ -\frac{1}{h_y}\frac{1}{2}(1-r) & -\frac{1}{h_y}\frac{1}{2}(1+r) & \frac{1}{h_y}\frac{1}{2}(1+r) & \frac{1}{h_y}\frac{1}{2}(1-r) \end{pmatrix}
$$

Then

$$
\boldsymbol{B}^T(r,s) \cdot \boldsymbol{B}(r,s) \tag{C.24}
$$

$$
= \begin{pmatrix} -\frac{1}{h_x}\frac{1}{2}(1-s) & -\frac{1}{h_y}\frac{1}{2}(1-r) \\ \frac{1}{h_x}\frac{1}{2}(1-s) & -\frac{1}{h_y}\frac{1}{2}(1+r) \\ \frac{1}{h_x}\frac{1}{2}(1+s) & \frac{1}{h_y}\frac{1}{2}(1+r) \\ -\frac{1}{h_x}\frac{1}{2}(1+s) & \frac{1}{h_y}\frac{1}{2}(1-r) \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{h_x}\frac{1}{2}(1-s) & \frac{1}{h_x}\frac{1}{2}(1-s) & \frac{1}{h_x}\frac{1}{2}(1+s) & -\frac{1}{h_x}\frac{1}{2}(1+s) \\ -\frac{1}{h_y}\frac{1}{2}(1-r) & -\frac{1}{h_y}\frac{1}{2}(1+r) & \frac{1}{h_y}\frac{1}{2}(1+r) & \frac{1}{h_y}\frac{1}{2}(1-r) \end{pmatrix} \tag{C.25}
$$

$$
= \frac{1}{4h_x^2} \begin{pmatrix} -(1-s) & -(1-r) \\ (1-s) & -(1+r) \\ (1+s) & (1+r) \\ -(1+s) & (1-r) \end{pmatrix} \cdot \begin{pmatrix} -(1-s) & (1-s) & (1+s) & -(1+s) \\ -(1-r) & -(1+r) & (1+r) & (1-r) \end{pmatrix} \tag{C.26}
$$

$$
= \frac{1}{4h_x^2} \begin{pmatrix} (1-r)^2 & (1-r^2) & -(1-r^2) & -(1-r)^2 \\ (1-r^2) & (1+r)^2 & -(1+r)^2 & -(1-r^2) \\ -(1-r^2) & -(1+r)^2 & (1+r)^2 & (1-r^2) \\ -(1-r)^2 & -(1-r^2) & (1-r^2) & (1-r)^2 \end{pmatrix} \tag{C.27}
$$

$$
= \frac{1}{4h_y^2} \begin{pmatrix} (1-s)^2 & -(1-s)^2 & -(1-s^2) & (1-s^2) \\ -(1-s)^2 & (1-s)^2 & (1-s^2) & -(1-s^2) \\ -(1-s^2) & (1-s^2) & (1+s)^2 & -(1+s)^2 \\ (1-s^2) & -(1-s^2) & -(1+s)^2 & (1+s)^2 \end{pmatrix} \tag{C.28}
$$

$$
\tag{C.29}
$$

So in the end

$$
\begin{aligned}
\boldsymbol{K}_d^e &= k\frac{h_x h_y}{4}\int_{-1}^{+1}\int_{-1}^{+1}\frac{1}{4h_x^2}\begin{pmatrix} (1-r)^2 & (1-r^2) & -(1-r^2) & -(1-r)^2 \\ (1-r^2) & (1+r)^2 & -(1+r)^2 & -(1-r^2) \\ -(1-r^2) & -(1+r)^2 & (1+r)^2 & (1-r^2) \\ -(1-r)^2 & -(1-r^2) & (1-r^2) & (1-r)^2 \end{pmatrix}\,drds \quad\text{(C.30)} \\[2mm]
&+ \frac{h_x h_y}{4}\int_{-1}^{+1}\int_{-1}^{+1}\frac{1}{4h_y^2}\begin{pmatrix} (1-s)^2 & -(1-s)^2 & -(1-s^2) & (1-s^2) \\ -(1-s)^2 & (1-s)^2 & (1-s^2) & -(1-s^2) \\ -(1-s^2) & (1-s^2) & (1+s)^2 & -(1+s)^2 \\ (1-s^2) & -(1-s^2) & -(1+s)^2 & (1+s)^2 \end{pmatrix}\,drds \quad\text{(C.31)} \\[2mm]
&= k\frac{kh_y}{8h_x}\int_{-1}^{+1}\begin{pmatrix} (1-r)^2 & (1-r^2) & -(1-r^2) & -(1-r)^2 \\ (1-r^2) & (1+r)^2 & -(1+r)^2 & -(1-r^2) \\ -(1-r^2) & -(1+r)^2 & (1+r)^2 & (1-r^2) \\ -(1-r)^2 & -(1-r^2) & (1-r^2) & (1-r)^2 \end{pmatrix}\,dr \quad\text{(C.32)} \\[2mm]
&+ \frac{kh_x}{8h_y}\int_{-1}^{+1}\begin{pmatrix} (1-s)^2 & -(1-s)^2 & -(1-s^2) & (1-s^2) \\ -(1-s)^2 & (1-s)^2 & (1-s^2) & -(1-s^2) \\ -(1-s^2) & (1-s^2) & (1+s)^2 & -(1+s)^2 \\ (1-s^2) & -(1-s^2) & -(1+s)^2 & (1+s)^2 \end{pmatrix}\,ds \quad\text{(C.33)} \\[2mm]
&= k\frac{kh_y}{8h_x}\frac{4}{3}\begin{pmatrix} 2 & 1 & -1 & 2 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{pmatrix} + \frac{kh_x}{8h_y}\frac{4}{3}\begin{pmatrix} 2 & -2 & -1 & 1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{pmatrix} \quad\text{(C.34)} \\[2mm]
&= \frac{kh_x h_y}{6}\begin{pmatrix} \frac{2}{h_x^2}+\frac{2}{h_y^2} & -\frac{2}{h_x^2}+\frac{1}{h_y^2} & -\frac{1}{h_x^2}-\frac{1}{h_y^2} & \frac{1}{h_x^2}-\frac{2}{h_y^2} \\ -\frac{2}{h_x^2}+\frac{1}{h_y^2} & \frac{2}{h_x^2}+\frac{2}{h_y^2} & \frac{1}{h_x^2}-\frac{2}{h_y^2} & -\frac{1}{h_x^2}-\frac{1}{h_y^2} \\ -\frac{1}{h_x^2}-\frac{1}{h_y^2} & \frac{1}{h_x^2}-\frac{2}{h_y^2} & \frac{2}{h_x^2}+\frac{2}{h_y^2} & -\frac{2}{h_x^2}+\frac{1}{h_y^2} \\ \frac{1}{h_x^2}-\frac{2}{h_y^2} & -\frac{1}{h_x^2}-\frac{1}{h_y^2} & -\frac{2}{h_x^2}+\frac{1}{h_y^2} & \frac{2}{h_x^2}+\frac{2}{h_y^2} \end{pmatrix} \quad\text{(C.35)}
\end{aligned}
$$

Advection matrix

$$
\boldsymbol{K}_a = \rho C_p \frac{h_x h_y}{4}\int_{-1}^{+1}\int_{-1}^{+1}\boldsymbol{N}^T(r,s)(\vec{\mathbf{v}}\cdot\boldsymbol{B}(r,s))\,drds
$$

with

$$
\vec{\mathbf{v}}\cdot\boldsymbol{B}(r,s) = \left(-\frac{u}{2h_x}(1-s)-\frac{v}{2h_y}(1-r) \quad \frac{u}{2h_x}(1-s)-\frac{v}{2h_y}(1+r) \quad \frac{u}{2h_x}(1+s)+\frac{v}{2h_y}(1+r) \quad -\frac{u}{2h_x}(1+s)+\right.
$$

1263

Assuming that the velocity is constant within the element (which is almost always not true!), we can write:

$$
\begin{aligned}
\boldsymbol{K}_a &= \rho C_p \frac{h_x h_y}{16} \frac{v}{2h_y} \int_{-1}^{+1} \int_{-1}^{+1} \begin{pmatrix} (1-r)(1-s) \\ (1+r)(1-s) \\ (1+r)(1+s) \\ (1-r)(1+s) \end{pmatrix} \begin{pmatrix} -(1-r) & -(1+r) & (1+r) & (1-r) \end{pmatrix} \, dr\,ds \quad (C.36) \\[2mm]
&+ \rho C_p \frac{h_x h_y}{16} \frac{u}{2h_x} \int_{-1}^{+1} \int_{-1}^{+1} \begin{pmatrix} (1-r)(1-s) \\ (1+r)(1-s) \\ (1+r)(1+s) \\ (1-r)(1+s) \end{pmatrix} \begin{pmatrix} -(1-s) & (1-s) & (1+s) & -(1+s) \end{pmatrix} \, dr\,ds \quad (C.37) \\[2mm]
&= \rho C_p \frac{h_x v}{32} \int_{-1}^{+1} \int_{-1}^{+1} \begin{pmatrix} (1-r)(1-s) \\ (1+r)(1-s) \\ (1+r)(1+s) \\ (1-r)(1+s) \end{pmatrix} \begin{pmatrix} -(1-r) & -(1+r) & (1+r) & (1-r) \end{pmatrix} \, dr\,ds \quad (C.38) \\[2mm]
&+ \rho C_p \frac{h_y u}{32} \int_{-1}^{+1} \int_{-1}^{+1} \begin{pmatrix} (1-r)(1-s) \\ (1+r)(1-s) \\ (1+r)(1+s) \\ (1-r)(1+s) \end{pmatrix} \begin{pmatrix} -(1-s) & (1-s) & (1+s) & -(1+s) \end{pmatrix} \, dr\,ds \quad (C.39) \\[2mm]
&= \rho C_p \frac{1}{12} \left(vh_x \begin{pmatrix} -2 & -1 & 1 & 2 \\ -1 & -2 & 2 & 1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{pmatrix} + uh_y \begin{pmatrix} -2 & 2 & 1 & -1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{pmatrix} \right) \quad (C.40)
\end{aligned}
$$

and finally

$$
\boldsymbol{K}_a = \frac{\rho C_p}{3} \begin{pmatrix}
-\frac{1}{2}uh_y - \frac{1}{2}vh_x & \frac{1}{2}uh_y - \frac{1}{4}vh_x & \frac{1}{4}uh_y + \frac{1}{4}vh_x & -\frac{1}{4}uh_y + \frac{1}{2}vh_x \\[2mm]
-\frac{1}{2}uh_y - \frac{1}{4}vh_x & \frac{1}{2}uh_y - \frac{1}{2}vh_x & \frac{1}{4}uh_y + \frac{1}{2}vh_x & -\frac{1}{4}uh_y + \frac{1}{4}vh_x \\[2mm]
-\frac{1}{4}uh_y - \frac{1}{4}vh_x & \frac{1}{4}uh_y - \frac{1}{2}vh_x & \frac{1}{2}uh_y + \frac{1}{2}vh_x & -\frac{1}{2}uh_y + \frac{1}{4}vh_x \\[2mm]
-\frac{1}{4}uh_y - \frac{1}{2}vh_x & \frac{1}{4}uh_y - \frac{1}{4}vh_x & \frac{1}{2}uh_y + \frac{1}{4}vh_x & -\frac{1}{2}uh_y + \frac{1}{2}vh_x
\end{pmatrix}
$$

Matrices for D.G.

In the context of Discontinuous Galerkin methods we will need

$$
\begin{aligned}
\boldsymbol{J}_x &= \int_\square \partial_x \vec{N}^T(x,y)\vec{N}(x,y)\,dxdy \\
\boldsymbol{J}_y &= \int_\square \partial_y \vec{N}^T(x,y)\vec{N}(x,y)\,dxdy
\end{aligned}
$$

$$ (C.41) $$

We have

$$
\partial_x N_i(x,y) = \frac{\partial N_i}{\partial r}\frac{\partial r}{\partial x} \qquad \text{and} \qquad \partial_y N_i(x,y) = \frac{\partial N_i}{\partial s}\frac{\partial s}{\partial y}
$$

Since

$$
r = \frac{2}{h_x}(x - x_0) - 1 \qquad \text{and} \qquad s = \frac{2}{h_y}(y - y_0) - 1
$$

then

$$
\frac{\partial r}{\partial x} = \frac{2}{h_x} \qquad \frac{\partial s}{\partial y} = \frac{2}{h_y}
$$

1264

so

$$\boldsymbol{J}_x = \int_\square \partial_x \vec{N}^T(x,y)\vec{N}(x,y)dxdy$$

$$= \frac{2}{h_x}\frac{h_x h_y}{4}\int_{-1}^{+1}\int_{-1}^{+1}\partial_r\vec{N}^T(r,s)\vec{N}(r,s)drds$$

$$= \frac{h_y}{2}\int_{-1}^{+1}\int_{-1}^{+1}\begin{pmatrix}-\frac{1}{4}(1-s)\\+\frac{1}{4}(1-s)\\+\frac{1}{4}(1+s)\\-\frac{1}{4}(1+s)\end{pmatrix}\begin{pmatrix}\frac{1}{4}(1-r)(1-s)&\frac{1}{4}(1+r)(1-s)&\frac{1}{4}(1+r)(1+s)&\frac{1}{4}(1-r)(1+s)\end{pmatrix}$$

$$= \frac{h_y}{32}\int_{-1}^{+1}\int_{-1}^{+1}\begin{pmatrix}-(1-r)(1-s)^2&-(1+r)(1-s)^2&-(1+r)(1-s^2)&-(1-r)(1-s^2)\\(1-r)(1-s)^2&(1+r)(1-s)^2&(1+r)(1-s^2)&(1-r)(1-s^2)\\(1-r)(1-s^2)&(1+r)(1-s^2)&(1+r)(1+s)^2&(1-r)(1+s)^2\\-(1-r)(1-s^2)&-(1+r)(1-s^2)&-(1+r)(1+s)^2&-(1-r)(1+s)^2\end{pmatrix}drds$$

$$= \frac{h_y}{32}\begin{pmatrix}-16/3&-16/3&-8/3&-8/3\\16/3&16/3&8/3&8/3\\8/3&8/3&16/3&16/3\\-8/3&-8/3&-16/3&-16/3\end{pmatrix}$$

$$= \frac{h_y}{12}\begin{pmatrix}-2&-2&-1&-1\\2&2&1&1\\1&1&2&2\\-1&-1&-2&-2\end{pmatrix}$$

$$\boldsymbol{J}_y = \int_\square \partial_y \vec{N}^T(x,y)\vec{N}(x,y)dxdy$$

$$= \frac{2}{h_y}\frac{h_x h_y}{4}\int_{-1}^{+1}\int_{-1}^{+1}\begin{pmatrix}-\frac{1}{4}(1-r)\\-\frac{1}{4}(1+r)\\+\frac{1}{4}(1+r)\\+\frac{1}{4}(1-r)\end{pmatrix}\begin{pmatrix}\frac{1}{4}(1-r)(1-s)&\frac{1}{4}(1+r)(1-s)&\frac{1}{4}(1+r)(1+s)&\frac{1}{4}(1-r)(1+s)\end{pmatrix}$$

$$= \frac{h_x}{32}\int_{-1}^{+1}\int_{-1}^{+1}\begin{pmatrix}-(1-r)^2(1-s)&-(1-r^2)(1-s)&-(1-r^2)(1+s)&-(1-r)^2(1+s)\\-(1-r^2)(1-s)&-(1+r)^2(1-s)&-(1+r)^2(1+s)&-(1-r^2)(1+s)\\(1-r^2)(1-s)&(1+r)^2(1-s)&(1+r)^2(1+s)&(1-r^2)(1+s)\\(1-r)^2(1-s)&(1-r^2)(1-s)&(1-r^2)(1+s)&(1-r)^2(1+s)\end{pmatrix}drds$$

$$= \frac{h_x}{32}\begin{pmatrix}-16/3&-8/3&-8/3&-16/3\\-8/3&-16/3&-16/3&-8/3\\8/3&16/3&16/3&8/3\\16/3&8/3&8/3&16/3\end{pmatrix}$$

$$= \frac{h_x}{12}\begin{pmatrix}-2&-1&-1&-2\\-1&-2&-2&-1\\1&2&2&1\\2&1&1&2\end{pmatrix}$$

Computing matrix C_1

$$C_1 = \int_{\partial\Omega_1} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma$$

The edge $\partial\Omega_1$ is bounded by the coordinates of nodes x_1, y_1 and x_2, y_2. This segment can be parameterised by $t \in [0,1]$:

$$\vec{r}(t) = (1-t)\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + t\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} (x_2 - x_1)t + x_1 \\ (y_2 - y_1)t + y_1 \end{pmatrix}$$

Let us assume that C is a smooth curve and that it is given by the parametric equations $x = h(t)$, $y = g(t)$ and $a \le t \le b$. The line integral of a function $f(x,y)$ over C is computed as follows.

$$\int_C f(x,y)ds = \int_a^b f(h(t), g(t))\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt$$

In our case $dx/dt = x_2 - x_1$ and $dy/dt = y_2 - y_1$ so

$$\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = h_x$$

Then

$$C_1 = \int_{\partial\Omega_1} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma$$

$$= \begin{pmatrix} \int_{\partial\Omega_1} N_1(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_1(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_1(x,y)N_3(x,y)d\Gamma & \int_{\partial\Omega_1} N_1(x,y)N_4(x,y) \\ \int_{\partial\Omega_1} N_2(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_2(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_2(x,y)N_3(x,y)d\Gamma & \int_{\partial\Omega_1} N_2(x,y)N_4(x,y) \\ \int_{\partial\Omega_1} N_3(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_3(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_3(x,y)N_3(x,y)d\Gamma & \int_{\partial\Omega_1} N_3(x,y)N_4(x,y) \\ \int_{\partial\Omega_1} N_4(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_4(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_4(x,y)N_3(x,y)d\Gamma & \int_{\partial\Omega_1} N_4(x,y)N_4(x,y) \end{pmatrix}$$

$$= h_x \begin{pmatrix} \int_0^1 N_1(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_1(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_1(x(t),y(t))N_3(x(t),y(\\ \int_0^1 N_2(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_2(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_2(x(t),y(t))N_3(x(t),y(\\ \int_0^1 N_3(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_3(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_3(x(t),y(t))N_3(x(t),y(\\ \int_0^1 N_4(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_4(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_4(x(t),y(t))N_3(x(t),y(\end{pmatrix}$$

On this edge $y_1 = y_2$ so $y(t) = y_1$.

$$
\begin{aligned}
N_1(x(t), y(t)) &= \frac{x_3 - x(t)}{x_3 - x_1} \frac{y_3 - y(t)}{y_3 - y_1} \\
&= \frac{x_3 - (x_2 - x_1)t - x_1}{x_3 - x_1} \frac{y_3 - y_1}{y_3 - y_1} \\
&= \frac{x_3 - x_1 - (x_3 - x_1)t}{x_3 - x_1} \frac{y_3 - y_1}{y_3 - y_1} \\
&= (1 - t) && \text{(C.44)} \\
N_2(x(t), y(t)) &= \frac{x(t) - x_1}{x_2 - x_1} \frac{y_3 - y(t)}{y_3 - y_2} \\
&= \frac{(x_2 - x_1)t + x_1 - x_1}{x_2 - x_1} \frac{y_3 - y_1}{y_3 - y_2} \\
&= t && \text{(C.45)} \\
N_3(x(t), y(t)) &= 0 \qquad \text{by construction on edge 1} && \text{(C.46)} \\
N_4(x(t), y(t)) &= 0 \qquad \text{by construction on edge 1} && \text{(C.47)}
\end{aligned}
$$

so that

$$
\begin{aligned}
\boldsymbol{C}_1 &= h_x \begin{pmatrix} \int_0^1 (1-t)^2 dt & \int_0^1 (1-t)t\, dt & 0 & 0 \\ \int_0^1 t(1-t)dt & \int_0^1 t^2 dt & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
&= h_x \begin{pmatrix} 1/3 & 1/6 & 0 & 0 \\ 1/6 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
&= \frac{h_x}{6} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} && \text{(C.48)}
\end{aligned}
$$

Computing matrix C_3

```
      edge3
  4-------3
  |       |
  |       |
  |       |
  1-------2
```
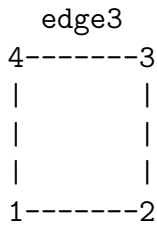
$$
\boldsymbol{C}_3 = \int_{\partial\Omega_3} \vec{N}^T(x, y)\vec{N}(x, y)d\Gamma = \int_{3 \to 4} \vec{N}^T(x, y)\vec{N}(x, y)d\Gamma
$$

The edge $\partial\Omega_3$ is bounded by the coordinates of nodes x_3, y_3 and x_4, y_4. This segment can be parameterised by $t \in [0, 1]$:

$$
\vec{r}(t) = (1 - t)\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} + t\begin{pmatrix} x_4 \\ y_4 \end{pmatrix} = \begin{pmatrix} (x_4 - x_3)t + x_3 \\ (y_4 - y_3)t + y_3 \end{pmatrix} = \begin{pmatrix} (x_4 - x_3)t + x_3 \\ y_3 \end{pmatrix}
$$

since $y_3 = y_4$. Here again the jacobian of the transformation is h_x.

$$
\begin{aligned}
N_1(x(t), y(t)) &= 0 \qquad \text{by construction on edge 3} \\
N_2(x(t), y(t)) &= 0 \qquad \text{by construction on edge 3} \\
N_3(x(t), y(t)) &= \frac{x(t) - x_1}{x_3 - x_1} \frac{y(t) - y_1}{y_3 - y_1} \\
&= \frac{(x_4 - x_3)t + x_3 - x_1}{x_3 - x_1} \frac{y_3 - y_1}{y_3 - y_1} \\
&= \frac{(x_4 - x_3)t + x_3 - x_1}{x_3 - x_1} \\
&= 1 - t \\
N_4(x(t), y(t)) &= \frac{x(t) - x_3}{x_4 - x_3} \frac{y(t) - y_1}{y_4 - y_1} \\
&= \frac{(x_4 - x_3)t + x_3 - x_3}{x_4 - x_3} \frac{y_3 - y_1}{y_4 - y_1} \\
&= t
\end{aligned}
\tag{C.49}
$$

Then

$$
\begin{aligned}
\boldsymbol{C}_3 &= h_x \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & \int_0^1 (1-t)^2 dt & \int_0^1 (1-t)t\, dt \\
0 & 0 & \int_0^1 t(1-t) dt & \int_0^1 t^2 dt
\end{pmatrix} \\
&= h_x \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1/3 & 1/6 \\
0 & 0 & 1/6 & 1/3
\end{pmatrix} \\
&= \frac{h_x}{6} \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 2 & 1 \\
0 & 0 & 1 & 2
\end{pmatrix}
\end{aligned}
\tag{C.50}
$$

The $\mathbb{K}$ and $\mathbb{G}$ matrices for $Q_1 \times P_0$ in 2D

Let us consider a regular grid composed of $nelx \times nely$ rectangular linear elements on a domain of dimensions $L_x \times L_y$. We are here interested in the elemental matrices $\mathbb{K}_e$ and $\mathbb{G}_e$. Borrowing (for example) from stone 1 we can write a simple code which compute these matrices by means of numerical integration. This code is available there: `python_codes/Gel/compute_K_G_S_q1p0.py`.

We start with square elements and a constant viscosity $\eta = 1$. We find that $\mathbb{K}_e$ is independent of the resolution (i.e. independent of $nelx = nely$):

$$
\mathbb{K}_e = \begin{pmatrix}
1 & 0.25 & -0.5 & -0.25 & -0.5 & -0.25 & 0 & 0.25 \\
0.25 & 1 & 0.25 & 0 & -0.25 & -0.5 & -0.25 & -0.5 \\
-0.5 & 0.25 & 1 & -0.25 & 0 & -0.25 & -0.5 & 0.25 \\
-0.25 & 0 & -0.25 & 1 & 0.25 & -0.5 & 0.25 & -0.5 \\
-0.5 & -0.25 & 0 & 0.25 & 1 & 0.25 & -0.5 & -0.25 \\
-0.25 & -0.5 & -0.25 & -0.5 & 0.25 & 1 & 0.25 & 0 \\
0 & -0.25 & -0.5 & 0.25 & -0.5 & 0.25 & 1 & -0.25 \\
0.25 & -0.5 & 0.25 & -0.5 & -0.25 & 0 & -0.25 & 1
\end{pmatrix}
$$

and its lumped version $\tilde{\mathbb{K}}_{i,j} = \sum_j |\mathbb{K}_{i,j}|$ is:

$$\tilde{\mathbb{K}}_e = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

Let us now turn to non-square elements. Let us print $\mathbb{K}_e$ and $\mathbb{G}_e$ for 2 resolutions, 4x8 and 8x4:

$$\mathbb{K}_e = \begin{pmatrix} 1 & 0.25 & 0 & -0.25 & -0.5 & -0.25 & -0.5 & 0.25 \\ 0.25 & 1.5 & 0.25 & 0.5 & -0.25 & -0.75 & -0.25 & -1.25 \\ 0 & 0.25 & 1 & -0.25 & -0.5 & -0.25 & -0.5 & 0.25 \\ -0.25 & 0.5 & -0.25 & 1.5 & 0.25 & -1.25 & 0.25 & -0.75 \\ -0.5 & -0.25 & -0.5 & 0.25 & 1 & 0.25 & 0 & -0.25 \\ -0.25 & -0.75 & -0.25 & -1.25 & 0.25 & 1.5 & 0.25 & 0.5 \\ -0.5 & -0.25 & -0.5 & 0.25 & 0 & 0.25 & 1 & -0.25 \\ 0.25 & -1.25 & 0.25 & -0.75 & -0.25 & 0.5 & -0.25 & 1.5 \end{pmatrix} \qquad \mathbb{G}_e = \begin{pmatrix} 0.0625 \\ 0.125 \\ -0.0625 \\ 0.125 \\ -0.0625 \\ -0.125 \\ 0.0625 \\ -0.125 \end{pmatrix}$$

$$\mathbb{K}_e = \begin{pmatrix} 1.5 & 0.25 & -1.25 & -0.25 & -0.75 & -0.25 & 0.5 & 0.25 \\ 0.25 & 1 & 0.25 & -0.5 & -0.25 & -0.5 & -0.25 & 0 \\ -1.25 & 0.25 & 1.5 & -0.25 & 0.5 & -0.25 & -0.75 & 0.25 \\ -0.25 & -0.5 & -0.25 & 1 & 0.25 & 0 & 0.25 & -0.5 \\ -0.75 & -0.25 & 0.5 & 0.25 & 1.5 & 0.25 & -1.25 & -0.25 \\ -0.25 & -0.5 & -0.25 & 0 & 0.25 & 1 & 0.25 & -0.5 \\ 0.5 & -0.25 & -0.75 & 0.25 & -1.25 & 0.25 & 1.5 & -0.25 \\ 0.25 & 0 & 0.25 & -0.5 & -0.25 & -0.5 & -0.25 & 1 \end{pmatrix} \qquad \mathbb{G}_e = \begin{pmatrix} 0.125 \\ 0.0625 \\ -0.125 \\ 0.0625 \\ -0.125 \\ -0.0625 \\ 0.125 \\ -0.0625 \end{pmatrix}$$

We find that both matrices K_e and $\mathbb{G}_e$ are different from each other and different from the ones obtained with square elements. We have no other choice than computing these by hand in order to express these as a function of h_x and h_y. Let us start with $\mathbb{K}_e$:

$$\mathbb{K}_e \;=\; \iint_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \; dV = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \boldsymbol{B}^T(x,y) \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}(x,y) \; dxdy \qquad \text{with} \qquad \boldsymbol{C}_\eta = \eta \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In a rectangle bounded by $[x_1, x_3] \times [y_1, y_3]$ the basis functions are given by:

$$\begin{aligned} \mathcal{N}_1^\gamma(x,y) &= \left(\frac{x_3 - x}{h_x}\right)\left(\frac{y_3 - y}{h_y}\right) \\ \mathcal{N}_2^\gamma(x,y) &= \left(\frac{x - x_1}{h_x}\right)\left(\frac{y_3 - y}{h_y}\right) \\ \mathcal{N}_3^\gamma(x,y) &= \left(\frac{x - x_1}{h_x}\right)\left(\frac{y - y_1}{h_y}\right) \\ \mathcal{N}_4^\gamma(x,y) &= \left(\frac{x_3 - x}{h_x}\right)\left(\frac{y - y_1}{h_y}\right) \end{aligned}$$

so that

$$\frac{\partial \mathcal{N}_1^\gamma}{\partial x} = -\frac{1}{h_x}\frac{y_3 - y}{h_y} = -\frac{1}{h_x}\frac{1}{2}(1 - s)$$

$$\frac{\partial \mathcal{N}_2^\gamma}{\partial x} = \frac{1}{h_x}\frac{y_3 - y}{h_y} = +\frac{1}{h_x}\frac{1}{2}(1 - s)$$

$$\frac{\partial \mathcal{N}_3^\gamma}{\partial x} = \frac{1}{h_x}\frac{y - y_1}{h_y} = +\frac{1}{h_x}\frac{1}{2}(1 + s)$$

$$\frac{\partial \mathcal{N}_4^\gamma}{\partial x} = -\frac{1}{h_x}\frac{y - y_1}{h_y} = -\frac{1}{h_x}\frac{1}{2}(1 + s)$$

$$\frac{\partial \mathcal{N}_1^\gamma}{\partial y} = -\frac{1}{h_y}\frac{x_3 - x}{h_x} = -\frac{1}{h_y}\frac{1}{2}(1 - r)$$

$$\frac{\partial \mathcal{N}_2^\gamma}{\partial y} = -\frac{1}{h_y}\frac{x - x_1}{h_x} = -\frac{1}{h_y}\frac{1}{2}(1 + r)$$

$$\frac{\partial \mathcal{N}_3^\gamma}{\partial y} = \frac{1}{h_y}\frac{x - x_1}{h_x} = +\frac{1}{h_y}\frac{1}{2}(1 + r)$$

$$\frac{\partial \mathcal{N}_4^\gamma}{\partial y} = \frac{1}{h_y}\frac{x_3 - x}{h_x} = +\frac{1}{h_y}\frac{1}{2}(1 - r)$$

The matrix $\boldsymbol{B}$ is given by

$$\boldsymbol{B}(x, y) = \begin{pmatrix} \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & 0 & \frac{\partial \mathcal{N}_2^\gamma}{\partial x} & 0 & \frac{\partial \mathcal{N}_3^\gamma}{\partial x} & 0 & \frac{\partial \mathcal{N}_4^\gamma}{\partial x} & 0 \\ 0 & \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & 0 & \frac{\partial \mathcal{N}_2^\gamma}{\partial y} & 0 & \frac{\partial \mathcal{N}_3^\gamma}{\partial y} & 0 & \frac{\partial \mathcal{N}_4^\gamma}{\partial y} \\ \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & \frac{\partial \mathcal{N}_2^\gamma}{\partial y} & \frac{\partial \mathcal{N}_2^\gamma}{\partial x} & \frac{\partial \mathcal{N}_3^\gamma}{\partial y} & \frac{\partial \mathcal{N}_3^\gamma}{\partial x} & \frac{\partial \mathcal{N}_4^\gamma}{\partial y} & \frac{\partial \mathcal{N}_4^\gamma}{\partial x} \end{pmatrix}$$

so that

$$\boldsymbol{C}_\eta \cdot \boldsymbol{B} = \eta \begin{pmatrix} 2\frac{\partial \mathcal{N}_1^\gamma}{\partial x} & 0 & 2\frac{\partial \mathcal{N}_2^\gamma}{\partial x} & 0 & 2\frac{\partial \mathcal{N}_3^\gamma}{\partial x} & 0 & 2\frac{\partial \mathcal{N}_4^\gamma}{\partial x} & 0 \\ 0 & 2\frac{\partial \mathcal{N}_1^\gamma}{\partial y} & 0 & 2\frac{\partial \mathcal{N}_2^\gamma}{\partial y} & 0 & 2\frac{\partial \mathcal{N}_3^\gamma}{\partial y} & 0 & 2\frac{\partial \mathcal{N}_4^\gamma}{\partial y} \\ \frac{\partial \mathcal{N}_1^\gamma}{\partial y} & \frac{\partial \mathcal{N}_1^\gamma}{\partial x} & \frac{\partial \mathcal{N}_2^\gamma}{\partial y} & \frac{\partial \mathcal{N}_2^\gamma}{\partial x} & \frac{\partial \mathcal{N}_3^\gamma}{\partial y} & \frac{\partial \mathcal{N}_3^\gamma}{\partial x} & \frac{\partial \mathcal{N}_4^\gamma}{\partial y} & \frac{\partial \mathcal{N}_4^\gamma}{\partial x} \end{pmatrix}$$

Let us start with the diagonal terms

$$\mathbb{K}_{11} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \right)_{11} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2 \left(\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \right)^2 + \left(\frac{\partial \mathcal{N}_1^\gamma}{\partial y} \right)^2 \right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[2 \left(-\frac{1}{h_x} \frac{1}{2} (1-s) \right)^2 + \left(-\frac{1}{h_y} \frac{1}{2} (1-r) \right)^2 \right] drds$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[\frac{1}{2h_x^2} (1-s)^2 + \frac{1}{4h_y^2} (1-r)^2 \right] drds$$

$$= \frac{h_x h_y}{4} \left[\frac{1}{h_x^2} \frac{8}{3} + \frac{1}{2h_y^2} \frac{8}{3} \right]$$

$$= \frac{1}{3} \left(\frac{2h_y}{h_x} + \frac{h_x}{h_y} \right)$$

$$\mathbb{K}_{22} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \right)_{22} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[\left(\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \right)^2 + 2 \left(\frac{\partial \mathcal{N}_1^\gamma}{\partial y} \right)^2 \right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[\left(-\frac{1}{h_x} \frac{1}{2} (1-s) \right)^2 + 2 \left(-\frac{1}{h_y} \frac{1}{2} (1-r) \right)^2 \right] drds$$

$$= \frac{h_x h_y}{4} \left[\frac{1}{2h_x^2} \frac{8}{3} + \frac{1}{h_y^2} \frac{8}{3} \right]$$

$$= \frac{1}{3} \left(\frac{h_y}{h_x} + \frac{2h_x}{h_y} \right)$$

$$\mathbb{K}_{33} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \right)_{33} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2 \left(\frac{\partial \mathcal{N}_2^\gamma}{\partial x} \right)^2 + \left(\frac{\partial \mathcal{N}_2^\gamma}{\partial y} \right)^2 \right] dxdy$$

$$= \mathbb{K}_{11}$$

$$\mathbb{K}_{44} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B} \right)_{44} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[\left(\frac{\partial \mathcal{N}_2^\gamma}{\partial x} \right)^2 + 2 \left(\frac{\partial \mathcal{N}_2^\gamma}{\partial y} \right)^2 \right] dxdy$$

$$= \mathbb{K}_{22}$$

$$\mathbb{K}_{55} = \mathbb{K}_{11}$$
$$\mathbb{K}_{66} = \mathbb{K}_{22}$$
$$\mathbb{K}_{77} = \mathbb{K}_{11}$$
$$\mathbb{K}_{88} = \mathbb{K}_{22}$$

Let us now focus on the first row of $\mathbb{K}$:

$$\mathbb{K}_{12} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{11} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \frac{\partial \mathcal{N}_1^\gamma}{\partial y}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-\frac{1}{h_x}\frac{1}{2}(1-s) \cdot -\frac{1}{h_y}\frac{1}{2}(1-r)\right] drds$$

$$= \frac{1}{16} \int_{-1}^{+1} \int_{-1}^{+1} (1-s)(1-r) drds$$

$$= \frac{1}{4}$$

$$\mathbb{K}_{13} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{13} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \frac{\partial \mathcal{N}_2^\gamma}{\partial x} + \frac{\partial \mathcal{N}_1^\gamma}{\partial y} \frac{\partial \mathcal{N}_2^\gamma}{\partial y}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-2\frac{1}{h_x}\frac{1}{2}(1-s) \cdot \frac{1}{h_x}\frac{1}{2}(1-s) - \frac{1}{h_y}\frac{1}{2}(1-r) \cdot -\frac{1}{h_y}\frac{1}{2}(1+r)\right] drds$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-\frac{1}{2h_x^2}(1-s)(1-s) + \frac{1}{4h_y^2}(1-r)(1+r)\right] drds$$

$$= \frac{h_x h_y}{4} \left[-\frac{1}{2h_x^2}2\frac{8}{3} + \frac{1}{4h_y^2}\frac{4}{3}2\right]$$

$$= \frac{1}{3}\left(-\frac{2h_y}{h_x} + \frac{h_x}{2h_y}\right)$$

$$\mathbb{K}_{14} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{11} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[\frac{\partial \mathcal{N}_2^\gamma}{\partial x} \frac{\partial \mathcal{N}_1^\gamma}{\partial y}\right] dxdy$$

$$= -\frac{1}{4}$$

$$\mathbb{K}_{15} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{15} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \frac{\partial \mathcal{N}_3^\gamma}{\partial x} + \frac{\partial \mathcal{N}_1^\gamma}{\partial y} \frac{\partial \mathcal{N}_3^\gamma}{\partial y}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-2\frac{1}{h_x}\frac{1}{2}(1-s) \cdot \frac{1}{h_x}\frac{1}{2}(1+s) - \frac{1}{h_y}\frac{1}{2}(1-r) \cdot \frac{1}{h_y}\frac{1}{2}(1+r)\right] drds$$

$$= \frac{h_x h_y}{4} \left[-\frac{1}{2h_x^2}\frac{4}{3}2 - \frac{1}{4h_y^2}\frac{4}{3}2\right]$$

$$= \frac{1}{3}\left(-\frac{h_y}{h_x} - \frac{h_x}{2h_y}\right)$$

$$\mathbb{K}_{16} = -\frac{1}{4}$$

$$\mathbb{K}_{17} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{17} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2\frac{\partial \mathcal{N}_1^\gamma}{\partial x} \frac{\partial \mathcal{N}_4^\gamma}{\partial x} + \frac{\partial \mathcal{N}_1^\gamma}{\partial y} \frac{\partial \mathcal{N}_4^\gamma}{\partial y}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-2\frac{1}{h_x}\frac{1}{2}(1-s) \cdot \frac{1}{h_x}\frac{1}{2}(1+s) - \frac{1}{h_y}\frac{1}{2}(1-r) \cdot \frac{1}{h_y}\frac{1}{2}(1-r)\right] drds$$

$$= \frac{h_x h_y}{4} \left[\frac{1}{2h_x^2}2\frac{4}{3} - \frac{1}{4h_y^2}\frac{8}{3}2\right]$$

And now the other rows:

$$\mathbb{K}_{23} = \frac{1}{4}$$

$$\mathbb{K}_{24} = -\mathbb{K}_{17}$$

$$\mathbb{K}_{25} = -\frac{1}{4}$$

$$\mathbb{K}_{26} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{26} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2\frac{\partial \mathcal{N}_1^\gamma}{\partial y}\frac{\partial \mathcal{N}_3^\gamma}{\partial y} + \frac{\partial \mathcal{N}_1^\gamma}{\partial x}\frac{\partial \mathcal{N}_3^\gamma}{\partial x}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-2\frac{1}{h_y}\frac{1}{2}(1-r) \cdot \frac{1}{h_y}\frac{1}{2}(1+r) - \frac{1}{h_x}\frac{1}{2}(1-s) \cdot \frac{1}{h_x}\frac{1}{2}(1+s)\right] drds$$

$$= \frac{h_x h_y}{4} \left[-\frac{1}{2h_y^2}\frac{4}{3}2 - \frac{1}{4h_x^2}2\frac{4}{3}\right]$$

$$= \frac{1}{3}\left(-\frac{h_x}{h_y} - \frac{h_y}{2h_x}\right)$$

$$\mathbb{K}_{27} = -\frac{1}{4}$$

$$\mathbb{K}_{28} = \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left(\boldsymbol{B}^T \cdot \boldsymbol{C}_\eta \cdot \boldsymbol{B}\right)_{17} dxdy$$

$$= \int_{x_1}^{x_3} \int_{y_1}^{y_3} \left[2\frac{\partial \mathcal{N}_1^\gamma}{\partial y}\frac{\partial \mathcal{N}_4^\gamma}{\partial y} + \frac{\partial \mathcal{N}_1^\gamma}{\partial x}\frac{\partial \mathcal{N}_4^\gamma}{\partial x}\right] dxdy$$

$$= \frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1} \left[-2\frac{1}{h_y}\frac{1}{2}(1-r) \cdot \frac{1}{h_y}\frac{1}{2}(1-r) - \frac{1}{h_x}\frac{1}{2}(1-s) \cdot -\frac{1}{h_x}\frac{1}{2}(1+s)\right] drds$$

$$= \frac{h_x h_y}{4} \left[-\frac{1}{2h_y^2}\frac{8}{3}2 + \frac{1}{4h_x^2}2\frac{4}{3}\right]$$

$$= \frac{1}{3}\left(-\frac{2h_x}{h_y} + \frac{h_y}{2h_x}\right)$$

$$\mathbb{K}_{34} = -\frac{1}{4}$$

$$\mathbb{K}_{35} = \mathbb{K}_{17}$$

$$\mathbb{K}_{36} = -\frac{1}{4}$$

$$\mathbb{K}_{37} = \mathbb{K}_{15}$$

$$\mathbb{K}_{38} = +\frac{1}{4}$$

$$\mathbb{K}_{45} = \frac{1}{4}$$

$$\mathbb{K}_{46} = \mathbb{K}_{28}$$

$$\mathbb{K}_{47} = \frac{1}{4}$$

$$\mathbb{K}_{48} = \mathbb{K}_{26}$$

$$\mathbb{K}_{56} = \frac{1}{4}$$

$$\mathbb{K}_{57} = \mathbb{K}_{13}$$

$$\mathbb{K}_{58} = -\frac{1}{4}$$

$$\mathbb{K}_{67} = \frac{1}{4}$$

$$\mathbb{K}_{68} = -\mathbb{K}_{17}$$

$$\mathbb{K}_{78} = -\frac{1}{4}$$

so that (matrix $\mathbb{K}_e$ is symmetric so only half is shown here)[2]:

$$\begin{pmatrix}
\frac{1}{3}\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}\right) & \frac{1}{4} & \frac{1}{3}\left(-\frac{2h_y}{h_x}+\frac{h_x}{2h_y}\right) & -\frac{1}{4} & \frac{1}{3}\left(-\frac{h_y}{h_x}-\frac{h_x}{2h_y}\right) & -\frac{1}{4} & \frac{1}{3}\left(\frac{h_y}{h_x}-\frac{h_x}{h_y}\right) & \frac{1}{4} \\[2mm]
\cdot & \frac{1}{3}\left(\frac{h_y}{h_x}+\frac{2h_x}{h_y}\right) & \frac{1}{4} & -\frac{1}{3}\left(\frac{h_y}{h_x}-\frac{h_x}{h_y}\right) & -\frac{1}{4} & \frac{1}{3}\left(-\frac{h_x}{h_y}-\frac{h_y}{2h_x}\right) & -\frac{1}{4} & \frac{1}{3}\left(-\frac{2h_x}{h_y}+\frac{h_y}{2h_x}\right) \\[2mm]
\cdot & \cdot & \frac{1}{3}\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}\right) & -\frac{1}{4} & \frac{1}{3}\left(\frac{h_y}{h_x}-\frac{h_x}{h_y}\right) & -\frac{1}{4} & \frac{1}{3}\left(-\frac{h_y}{h_x}-\frac{h_x}{2h_y}\right) & \frac{1}{4} \\[2mm]
\cdot & \cdot & \cdot & \frac{1}{3}\left(\frac{h_y}{h_x}+\frac{2h_x}{h_y}\right) & \frac{1}{4} & \frac{1}{3}\left(-\frac{2h_x}{h_y}+\frac{h_y}{2h_x}\right) & \frac{1}{4} & \frac{1}{3}\left(-\frac{h_x}{h_y}-\frac{h_y}{2h_x}\right) \\[2mm]
\cdot & \cdot & \cdot & \cdot & \frac{1}{3}\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}\right) & \frac{1}{4} & \frac{1}{3}\left(-\frac{2h_y}{h_x}+\frac{h_x}{2h_y}\right) & -\frac{1}{4} \\[2mm]
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{3}\left(\frac{h_y}{h_x}+\frac{2h_x}{h_y}\right) & \frac{1}{4} & -\frac{1}{3}\left(\frac{h_y}{h_x}-\frac{h_x}{h_y}\right) \\[2mm]
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{3}\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}\right) & -\frac{1}{4} \\[2mm]
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{3}\left(\frac{h_y}{h_x}+\frac{2h_x}{h_y}\right)
\end{pmatrix}$$

For reference I present here under 3 matrices obtained with three coarse resolutions:

- 4×4 mesh

```
[[ 1.0000  0.2500 -0.5000 -0.2500 -0.5000 -0.2500  0.0000  0.2500]
 [ 0.2500  1.0000  0.2500 -0.0000 -0.2500 -0.5000 -0.2500 -0.5000]
 [-0.5000  0.2500  1.0000 -0.2500  0.0000 -0.2500 -0.5000  0.2500]
 [-0.2500 -0.0000 -0.2500  1.0000  0.2500 -0.5000  0.2500 -0.5000]
 [-0.5000 -0.2500  0.0000  0.2500  1.0000  0.2500 -0.5000 -0.2500]
 [-0.2500 -0.5000 -0.2500 -0.5000  0.2500  1.0000  0.2500 -0.0000]
 [ 0.0000 -0.2500 -0.5000  0.2500 -0.5000  0.2500  1.0000 -0.2500]
 [ 0.2500 -0.5000  0.2500 -0.5000 -0.2500 -0.0000 -0.2500  1.0000]]
```

- 7×5 mesh

```
[[ 1.1714  0.2500 -0.8143 -0.2500 -0.5857 -0.2500  0.2286  0.2500]
 [ 0.2500  0.9429  0.2500 -0.2286 -0.2500 -0.4714 -0.2500 -0.2429]
 [-0.8143  0.2500  1.1714 -0.2500  0.2286 -0.2500 -0.5857  0.2500]
 [-0.2500 -0.2286 -0.2500  0.9429  0.2500 -0.2429  0.2500 -0.4714]
 [-0.5857 -0.2500  0.2286  0.2500  1.1714  0.2500 -0.8143 -0.2500]
 [-0.2500 -0.4714 -0.2500 -0.2429  0.2500  0.9429  0.2500 -0.2286]
 [ 0.2286 -0.2500 -0.5857  0.2500 -0.8143  0.2500  1.1714 -0.2500]
 [ 0.2500 -0.2429  0.2500 -0.4714 -0.2500 -0.2286 -0.2500  0.9429]]
```

- 5×7 mesh

```
[[ 0.9429  0.2500 -0.2429 -0.2500 -0.4714 -0.2500 -0.2286  0.2500]
 [ 0.2500  1.1714  0.2500  0.2286 -0.2500 -0.5857 -0.2500 -0.8143]
 [-0.2429  0.2500  0.9429 -0.2500 -0.2286 -0.2500 -0.4714  0.2500]
 [-0.2500  0.2286 -0.2500  1.1714  0.2500 -0.8143  0.2500 -0.5857]
 [-0.4714 -0.2500 -0.2286  0.2500  0.9429  0.2500 -0.2429 -0.2500]
 [-0.2500 -0.5857 -0.2500 -0.8143  0.2500  1.1714  0.2500  0.2286]
 [-0.2286 -0.2500 -0.4714  0.2500 -0.2429  0.2500  0.9429 -0.2500]
 [ 0.2500 -0.8143  0.2500 -0.5857 -0.2500  0.2286 -0.2500  1.1714]]
```

---

[2]Values above are fully checked, values in matrix below should be re-checked to be sure

Turning now to the lumped version of $\mathbb{K}_e$, and because the sum of $\pm 1/4$ terms always add up to 1, we find:

$$
\begin{aligned}
\tilde{\mathbb{K}}_{11} &= |\mathbb{K}_{11}| + |\mathbb{K}_{13}| + |\mathbb{K}_{15}| + |\mathbb{K}_{17}| + 1 \\
\tilde{\mathbb{K}}_{22} &= |\mathbb{K}_{22}| + |\mathbb{K}_{24}| + |\mathbb{K}_{26}| + |\mathbb{K}_{28}| + 1 \\
\tilde{\mathbb{K}}_{33} &= |\mathbb{K}_{31}| + |\mathbb{K}_{33}| + |\mathbb{K}_{35}| + |\mathbb{K}_{37}| + 1 \\
\tilde{\mathbb{K}}_{44} &= |\mathbb{K}_{42}| + |\mathbb{K}_{44}| + |\mathbb{K}_{46}| + |\mathbb{K}_{48}| + 1 \\
\tilde{\mathbb{K}}_{55} &= |\mathbb{K}_{51}| + |\mathbb{K}_{53}| + |\mathbb{K}_{55}| + |\mathbb{K}_{57}| + 1 \\
\tilde{\mathbb{K}}_{66} &= |\mathbb{K}_{62}| + |\mathbb{K}_{64}| + |\mathbb{K}_{66}| + |\mathbb{K}_{68}| + 1 \\
\tilde{\mathbb{K}}_{77} &= |\mathbb{K}_{71}| + |\mathbb{K}_{73}| + |\mathbb{K}_{75}| + |\mathbb{K}_{77}| + 1 \\
\tilde{\mathbb{K}}_{88} &= |\mathbb{K}_{82}| + |\mathbb{K}_{84}| + |\mathbb{K}_{86}| + |\mathbb{K}_{88}| + 1
\end{aligned}
$$

Turning now to the $\mathbb{G}_e$ matrix, we find:

$$
\mathbb{G}_e = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{\mathcal{N}}^p \, dV = -\int_{\Omega_e}
\begin{pmatrix}
\frac{\partial \mathcal{N}_1}{\partial x} \\
\frac{\partial \mathcal{N}_1}{\partial y} \\
\frac{\partial \mathcal{N}_2}{\partial x} \\
\frac{\partial \mathcal{N}_2}{\partial y} \\
\frac{\partial \mathcal{N}_3}{\partial x} \\
\frac{\partial \mathcal{N}_3}{\partial y} \\
\frac{\partial \mathcal{N}_4}{\partial x} \\
\frac{\partial \mathcal{N}_4}{\partial y}
\end{pmatrix}
dV = -\frac{h_x h_y}{4} \int_{-1}^{+1} \int_{-1}^{+1}
\begin{pmatrix}
-\frac{1}{h_x}\frac{1}{2}(1-s) \\
-\frac{1}{h_y}\frac{1}{2}(1-r) \\
+\frac{1}{h_x}\frac{1}{2}(1-s) \\
-\frac{1}{h_y}\frac{1}{2}(1+r) \\
+\frac{1}{h_x}\frac{1}{2}(1+s) \\
+\frac{1}{h_y}\frac{1}{2}(1+r) \\
-\frac{1}{h_x}\frac{1}{2}(1+s) \\
+\frac{1}{h_y}\frac{1}{2}(1-r)
\end{pmatrix}
dr\,ds =
\begin{pmatrix}
h_y/2 \\
h_x/2 \\
-h_y/2 \\
+h_x/2 \\
-h_y/2 \\
-h_x/2 \\
+h_y/2 \\
-h_x/2
\end{pmatrix}
$$

which is Eq. (3.65) in Elman, Silvester, and Wathen [394]. Since the pressure is constant inside each element, then $\mathbb{G}_{el}$ is $(ndof_V * m_V, m_P) = (8 \times 1)$.

C.0.3 Quadrilaterals: rectangular quadratic elements
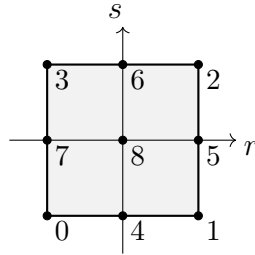
$$
\vec{\mathcal{N}}(r,s) =
\begin{pmatrix}
\mathcal{N}_1(r,s) \\
\mathcal{N}_2(r,s) \\
\mathcal{N}_3(r,s) \\
\mathcal{N}_4(r,s) \\
\mathcal{N}_5(r,s) \\
\mathcal{N}_6(r,s) \\
\mathcal{N}_7(r,s) \\
\mathcal{N}_8(r,s) \\
\mathcal{N}_9(r,s)
\end{pmatrix}
=
\begin{pmatrix}
\mathcal{N}_1(r)\mathcal{N}_1(s) \\
\mathcal{N}_2(r)\mathcal{N}_1(s) \\
\mathcal{N}_3(r)\mathcal{N}_1(s) \\
\mathcal{N}_1(r)\mathcal{N}_2(s) \\
\mathcal{N}_2(r)\mathcal{N}_2(s) \\
\mathcal{N}_3(r)\mathcal{N}_2(s) \\
\mathcal{N}_1(r)\mathcal{N}_3(s) \\
\mathcal{N}_2(r)\mathcal{N}_3(s) \\
\mathcal{N}_3(r)\mathcal{N}_3(s)
\end{pmatrix}
=
\begin{pmatrix}
\frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \\
(1-r^2)\frac{1}{2}s(s-1) \\
\frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \\
\frac{1}{2}r(r-1)(1-s^2) \\
(1-r^2)(1-s^2)\frac{1}{2}r(r+1)(1-s^2) \\
\frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \\
(1-r^2)\frac{1}{2}s(s+1) \\
\frac{1}{2}r(r+1)\frac{1}{2}s(s+1)
\end{pmatrix}
$$

The mass matrix on the reference element is then

$$
\boldsymbol{M}^e = \iint_\square
\begin{pmatrix}
\mathcal{N}_1\mathcal{N}_1 & \mathcal{N}_1\mathcal{N}_2 & \mathcal{N}_1\mathcal{N}_3 & \mathcal{N}_1\mathcal{N}_4 & \mathcal{N}_1\mathcal{N}_5 & \mathcal{N}_1\mathcal{N}_6 & \mathcal{N}_1\mathcal{N}_7 & \mathcal{N}_1\mathcal{N}_8 & \mathcal{N}_1\mathcal{N}_9 \\
\mathcal{N}_2\mathcal{N}_1 & \mathcal{N}_2\mathcal{N}_2 & \mathcal{N}_2\mathcal{N}_3 & \mathcal{N}_2\mathcal{N}_4 & \mathcal{N}_2\mathcal{N}_5 & \mathcal{N}_2\mathcal{N}_6 & \mathcal{N}_2\mathcal{N}_7 & \mathcal{N}_2\mathcal{N}_8 & \mathcal{N}_2\mathcal{N}_9 \\
\cdots \\
\mathcal{N}_9\mathcal{N}_1 & \mathcal{N}_9\mathcal{N}_2 & \mathcal{N}_9\mathcal{N}_3 & \mathcal{N}_9\mathcal{N}_4 & \mathcal{N}_9\mathcal{N}_5 & \mathcal{N}_9\mathcal{N}_6 & \mathcal{N}_9\mathcal{N}_7 & \mathcal{N}_9\mathcal{N}_8 & \mathcal{N}_9\mathcal{N}_9
\end{pmatrix}
drds
$$

with for example[3]

$$\iint_\square \mathcal{N}_1(r,s)\mathcal{N}_1(r,s)\ drds \quad = \quad \frac{16}{225} \tag{C.51}$$

$$\iint_\square \mathcal{N}_1(r,s)\mathcal{N}_2(r,s)\ drds \quad = \quad \frac{8}{225} \tag{C.52}$$

$$\iint_\square \mathcal{N}_1(r,s)\mathcal{N}_3(r,s)\ drds \quad = \quad -\frac{4}{225} \tag{C.53}$$

$$\iint_\square \mathcal{N}_1(r,s)\mathcal{N}_8(r,s)\ drds \quad = \quad \frac{-2}{225} \tag{C.54}$$

$$\iint_\square \mathcal{N}_1(r,s)\mathcal{N}_9(r,s)\ drds \quad = \quad \frac{1}{225} \tag{C.55}$$

$$\iint_\square \mathcal{N}_5(r,s)\mathcal{N}_5(r,s)\ drds \quad = \quad \frac{256}{225} \tag{C.56}$$

$$\tag{C.57}$$

In STONE 107, we find for a 3x2 mesh on domain 6x4:

```
M =
[[ 16.    8.   -4.    8.    4.   -2.   -4.   -2.    1.]
 [  8.   64.    8.    4.   32.    4.   -2.  -16.   -2.]
 [ -4.    8.   16.   -2.    4.    8.    1.   -2.   -4.]
 [  8.    4.   -2.   64.   32.  -16.    8.    4.   -2.]
 [  4.   32.    4.   32.  256.   32.    4.   32.    4.]
 [ -2.    4.    8.  -16.   32.   64.   -2.    4.    8.]
 [ -4.   -2.    1.    8.    4.   -2.   16.    8.   -4.]
 [ -2.  -16.   -2.    4.   32.    4.    8.   64.    8.]
 [  1.   -2.   -4.   -2.    4.    8.   -4.    8.   16.]] / 225
```

---

[3]Thank you WolframAlpha again!

C.0.4 Hexahedra: cuboid elements

We here assume that each element is a cuboid[4]. We set the domain size to $L_x = 4$, $L_x = 3$ and $L_z = 2$, with $nelx = 4$, $nely = 3$ and $nelz = 2$. Here again the viscosity is set to $\eta = 1$ so that we find that

$$\mathbb{K}_{el} = \frac{1}{8 \cdot 9}
\begin{pmatrix}
32 & 6 & 6 & -8 & -6 & -6 & -10 & -6 & -3 & 4 & 6 & 3 & 4 & 3 & 6 & -10 & -3 & -6 \\
6 & 32 & 6 & 6 & 4 & 3 & -6 & -10 & -3 & -6 & -8 & -6 & 3 & 4 & 6 & 3 & -4 & 3 \\
6 & 6 & 32 & 6 & 3 & 4 & 3 & 3 & -4 & 3 & 6 & 4 & -6 & -6 & -8 & -6 & -3 & -10 \\
-8 & 6 & 6 & 32 & -6 & -6 & 4 & -6 & -3 & -10 & 6 & 3 & -10 & 3 & 6 & 4 & -3 & -6 \\
-6 & 4 & 3 & -6 & 32 & 6 & 6 & -8 & -6 & 6 & -10 & -3 & -3 & -4 & 3 & -3 & 4 & 6 \\
-6 & 3 & 4 & -6 & 6 & 32 & -3 & 6 & 4 & -3 & 3 & -4 & 6 & -3 & -10 & 6 & -6 & -8 \\
10 & -6 & 3 & 4 & 6 & -3 & 32 & 6 & -6 & -8 & -6 & 6 & -8 & -3 & 3 & -4 & 3 & -3 \\
-6 & -10 & 3 & -6 & -8 & 6 & 6 & 32 & -6 & 6 & 4 & -3 & -3 & -8 & 3 & -3 & -10 & 6 \\
-3 & -3 & -4 & -3 & -6 & 4 & -6 & -6 & 32 & -6 & -3 & 4 & 3 & 3 & -8 & 3 & 6 & -10 \\
4 & -6 & 3 & -10 & 6 & -3 & -8 & 6 & -6 & 32 & -6 & 6 & -4 & -3 & 3 & -8 & 3 & -3 \\
6 & -8 & 6 & 6 & -10 & 3 & -6 & 4 & -3 & -6 & 32 & -6 & 3 & -10 & 6 & 3 & -8 & 3 \\
3 & -6 & 4 & 3 & -3 & -4 & 6 & -3 & 4 & 6 & -6 & 32 & -3 & 6 & -10 & -3 & 3 & -8 \\
4 & 3 & -6 & -10 & -3 & 6 & -8 & -3 & 3 & -4 & 3 & -3 & 32 & 6 & -6 & -8 & -6 & 6 \\
3 & 4 & -6 & 3 & -4 & -3 & -3 & -8 & 3 & -3 & -10 & 6 & 6 & 32 & -6 & 6 & 4 & -3 \\
6 & 6 & -8 & 6 & 3 & -10 & 3 & 3 & -8 & 3 & 6 & -10 & -6 & -6 & 32 & -6 & -3 & 4 \\
10 & 3 & -6 & 4 & -3 & 6 & -4 & -3 & 3 & -8 & 3 & -3 & -8 & 6 & -6 & 32 & -6 & 6 \\
-3 & -4 & -3 & -3 & 4 & -6 & 3 & -10 & 6 & 3 & -8 & 3 & -6 & 4 & -3 & -6 & 32 & -6 \\
-6 & 3 & -10 & -6 & 6 & -8 & -3 & 6 & -10 & -3 & 3 & -8 & 6 & -3 & 4 & 6 & -6 & 32 \\
-8 & -3 & -3 & -4 & 3 & 3 & 4 & 3 & 6 & -10 & -3 & -6 & -10 & -6 & -3 & 4 & 6 & 3 \\
-3 & -8 & -3 & -3 & -10 & -6 & 3 & 4 & 6 & 3 & -4 & 3 & -6 & -10 & -3 & -6 & -8 & -6 \\
-3 & -3 & -8 & -3 & -6 & -10 & -6 & -6 & -8 & -6 & -3 & -10 & 3 & 3 & -4 & 3 & 6 & 4 \\
-4 & -3 & -3 & -8 & 3 & 3 & -10 & 3 & 6 & 4 & -3 & -6 & 4 & -6 & -3 & -10 & 6 & 3 \\
3 & -10 & -6 & 3 & -8 & -3 & -3 & -4 & 3 & -3 & 4 & 6 & 6 & -8 & -6 & 6 & -10 & -3 \\
3 & -6 & -10 & 37-3 & -8 & 6 & -3 & -10 & 6 & -6 & -8 & -3 & 6 & 4 & -3 & 3 & -4 & -6
\end{pmatrix}$$

and

$$\mathbb{G}_{el} = \frac{1}{2 \cdot 9}
\begin{pmatrix}
1 \\ 1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ -1 \\ -1
\end{pmatrix}$$

---

[4] https://en.wikipedia.org/wiki/Cuboid

C.0.5 Triangles: linear elements

We start from the linear basis functions in the reference triangle given as a function of r, s:

$$
\begin{aligned}
\mathcal{N}_1(r, s) &= 1 - r - s & \text{(C.58)} \\
\mathcal{N}_2(r, s) &= r & \text{(C.59)} \\
\mathcal{N}_3(r, s) &= s & \text{(C.60)}
\end{aligned}
$$

and their derivatives:

$$
\begin{aligned}
\partial_r \mathcal{N}_1(r, s) &= -1 \\
\partial_r \mathcal{N}_2(r, s) &= 1 \\
\partial_r \mathcal{N}_3(r, s) &= 0 \\
\partial_s \mathcal{N}_1(r, s) &= -1 \\
\partial_s \mathcal{N}_2(r, s) &= 0 \\
\partial_s \mathcal{N}_3(r, s) &= 1
\end{aligned}
$$

We wish to compute the integral of a function $f(x, y)$ over the triangle by means of a change of variables $(x, y) \rightarrow (r, s)$:

$$
\begin{aligned}
\iint f(x, y) dx dy &= \iint f(x(r, s), y(r, s)) \left| \frac{\partial(x, y)}{\partial(r, s)} \right| dr ds \\
&= \iint f(x(r, s), y(r, s)) \begin{vmatrix} \partial x / \partial r & \partial x / \partial s \\ \partial y / \partial r & \partial y / \partial s \end{vmatrix} dr ds & \text{(C.61)}
\end{aligned}
$$

From

$$
x(r, s) = \sum_{i=1}^{3} \mathcal{N}_i(r, s) x_i \qquad \text{and} \qquad y(r, s) = \sum_{i=1}^{3} \mathcal{N}_i(r, s) y_i
$$

we can write

$$
\begin{aligned}
\frac{\partial x}{\partial r}(r, s) &= \sum_{i=1}^{3} \frac{\partial \mathcal{N}_i}{\partial r} x_i = \frac{\partial \mathcal{N}_1}{\partial r} x_1 + \frac{\partial \mathcal{N}_2}{\partial r} x_2 + \frac{\partial \mathcal{N}_3}{\partial r} x_3 = -x_1 + x_2 \\
\frac{\partial x}{\partial s}(r, s) &= \sum_{i=1}^{3} \frac{\partial \mathcal{N}_i}{\partial s} x_i = \frac{\partial \mathcal{N}_1}{\partial s} x_1 + \frac{\partial \mathcal{N}_2}{\partial s} x_2 + \frac{\partial \mathcal{N}_3}{\partial s} x_3 = -x_1 + x_3 \\
\frac{\partial y}{\partial r}(r, s) &= \sum_{i=1}^{3} \frac{\partial \mathcal{N}_i}{\partial r} y_i = \frac{\partial \mathcal{N}_1}{\partial r} x_1 + \frac{\partial \mathcal{N}_2}{\partial r} x_2 + \frac{\partial \mathcal{N}_3}{\partial r} y_3 = -y_1 + y_2 \\
\frac{\partial y}{\partial s}(r, s) &= \sum_{i=1}^{3} \frac{\partial \mathcal{N}_i}{\partial s} y_i = \frac{\partial \mathcal{N}_1}{\partial s} y_1 + \frac{\partial \mathcal{N}_2}{\partial s} y_2 + \frac{\partial \mathcal{N}_3}{\partial s} y_3 = -y_1 + y_3 & \text{(C.62)}
\end{aligned}
$$

Then

$$
\begin{vmatrix} \partial x / \partial r & \partial x / \partial s \\ \partial y / \partial r & \partial y / \partial s \end{vmatrix} = \begin{vmatrix} -x_1 + x_2 & -x_1 + x_3 \\ -y_1 + y_2 & -y_1 + y_3 \end{vmatrix} = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1) = 2S
$$

where S is the area of the triangle and which is independent of (r, s). Looking at the reference element, we find that when r goes from 0 to 1, s can only take values between 0 and $1 - r$.
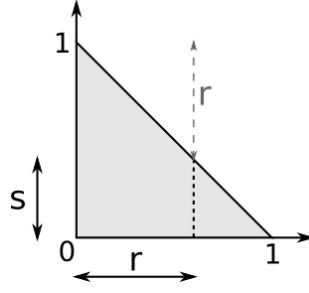
Then the bounds of the integrals are simply:

$$\iint_{\triangle} f(x,y)dxdy \;=\; 2S \int_0^1 \left(\int_0^{1-r} f(x(r,s), y(r,s))ds \right) dr \qquad (C.63)$$

and the mass matrix is given by

$$\boldsymbol{M}_e \;=\; 2S \int_0^1 \left[\int_0^{1-r} \begin{pmatrix} (1-r-s)^2 & (1-r-s)r & (1-r-s)s \\ (1-r-s)r & r^2 & rs \\ (1-r-s)s & rs & s^2 \end{pmatrix} ds \right] dr \qquad (C.64)$$

$$=\; 2S \int_0^1 \begin{pmatrix} \int_0^{1-r}(1-r-s)^2 ds & \int_0^{1-r}(1-r-s)rds & \int_0^{1-r}(1-r-s)sds \\[2mm] \int_0^{1-r}(1-r-s)rds & \int_0^{1-r} r^2 ds & \int_0^{1-r} rsds \\[2mm] \int_0^{1-r}(1-r-s)sds & \int_0^{1-r} rsds & \int_0^{1-r} s^2 ds \end{pmatrix} dr \quad (C.65)$$

$$=\; 2S \begin{pmatrix} 1/12 & 1/24 & 1/24 \\ 1/24 & 1/12 & 1/24 \\ 1/24 & 1/24 & 1/12 \end{pmatrix} \qquad (C.66)$$

$$=\; \frac{S}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \qquad (C.67)$$

This is Eq.(4.10e) of Li [825]. Also note that in the context of the heat transport equation this matrix is multiplied by ρC_p.

We will then compute the $\boldsymbol{J}_x$ and $\boldsymbol{J}_y$ matrices[5]. We start from the basis functions expressed in the (x,y) coordinate system:

$$\mathcal{N}_1(x,y) \;=\; \frac{1}{2S} \left(x_2 y_3 - x_3 y_2 + (y_2 - y_3)x + (x_3 - x_2)y \right)$$

$$\mathcal{N}_2(x,y) \;=\; \frac{1}{2S} \left(x_3 y_1 - x_1 y_3 + (y_3 - y_1)x + (x_1 - x_3)y \right)$$

$$\mathcal{N}_3(x,y) \;=\; \frac{1}{2S} \left(x_1 y_2 - x_2 y_1 + (y_1 - y_2)x + (x_2 - x_1)y \right)$$

---

[5]Nobody knows what these matrices are for ... needs to be lookied into.

where S is the area of the element. We then have

$$\partial_x \mathcal{N}_1(x,y) = \frac{1}{2S}(y_2 - y_3)$$
$$\partial_x \mathcal{N}_2(x,y) = \frac{1}{2S}(y_3 - y_1)$$
$$\partial_x \mathcal{N}_3(x,y) = \frac{1}{2S}(y_1 - y_2)$$
$$\partial_y \mathcal{N}_1(x,y) = \frac{1}{2S}(x_3 - x_2)$$
$$\partial_y \mathcal{N}_2(x,y) = \frac{1}{2S}(x_1 - x_3)$$
$$\partial_y \mathcal{N}_3(x,y) = \frac{1}{2S}(x_2 - x_1)$$

We start with

$$
\begin{aligned}
\boldsymbol{J}_x &= \iint_{\triangle} \partial_x \vec{\mathcal{N}}^T \vec{\mathcal{N}} dV \\
&= \iint_{\triangle} \begin{pmatrix} \frac{1}{2S}(y_2 - y_3) \\ \frac{1}{2S}(y_3 - y_1) \\ \frac{1}{2S}(y_1 - y_2) \end{pmatrix} \begin{pmatrix} \mathcal{N}_1(x,y) & \mathcal{N}_2(x,y) & \mathcal{N}_3(x,y) \end{pmatrix} dxdy \\
&= \frac{1}{2S} \begin{pmatrix} y_{23} \iint_{\triangle} \mathcal{N}_1 dxdy & y_{23} \iint_{\triangle} \mathcal{N}_2 dxdy & y_{23} \iint_{\triangle} \mathcal{N}_3 dxdy \\ y_{31} \iint_{\triangle} \mathcal{N}_1 dxdy & y_{31} \iint_{\triangle} \mathcal{N}_2 dxdy & y_{31} \iint_{\triangle} \mathcal{N}_3 dxdy \\ y_{12} \iint_{\triangle} \mathcal{N}_1 dxdy & y_{12} \iint_{\triangle} \mathcal{N}_2 dxdy & y_{12} \iint_{\triangle} \mathcal{N}_3 dxdy \end{pmatrix}
\end{aligned} \tag{C.68}
$$

where we have introduced the notation $x_{ij} = x_i - x_j$. We then need to compute

$$
\begin{aligned}
\iint_{\triangle} \mathcal{N}_1(x,y) dxdy &= 2S \int_0^1 \left(\int_0^{1-r} \mathcal{N}_1(x(r,s), y(r,s)) ds \right) dr \\
&= 2S \int_0^1 \left(\int_0^{1-r} (1 - r - s) ds \right) dr \\
&= 2S \frac{1}{6}
\end{aligned} \tag{C.69}
$$

$$
\begin{aligned}
\iint_{\triangle} \mathcal{N}_2(x,y) dxdy &= 2S \int_0^1 \left(\int_0^{1-r} \mathcal{N}_2(x(r,s), y(r,s)) ds \right) dr \\
&= 2S \int_0^1 \left(\int_0^{1-r} r ds \right) dr \\
&= 2S \frac{1}{6}
\end{aligned} \tag{C.70}
$$

$$
\begin{aligned}
\iint_{\triangle} \mathcal{N}_3(x,y) dxdy &= 2S \int_0^1 \left(\int_0^{1-r} \mathcal{N}_3(x(r,s), y(r,s)) ds \right) dr \\
&= 2S \int_0^1 \left(\int_0^{1-r} s ds \right) dr \\
&= 2S \frac{1}{6}
\end{aligned} \tag{C.71}
$$

verify!! FOR REAL 2025

Finally:

$$
\boldsymbol{J}_x = \frac{1}{6} \begin{pmatrix} y_{23} & y_{23} & y_{23} \\ y_{31} & y_{31} & y_{31} \\ y_{12} & y_{12} & y_{12} \end{pmatrix}
$$

Likewise

$$\boldsymbol{J}_y = \iint_{\triangle} \partial_y \vec{\mathcal{N}}^T \vec{\mathcal{N}} dV$$

$$= \iint_{\triangle} \left(\begin{array}{c} \frac{1}{2S}(x_3 - x_2) \\ \frac{1}{2S}(x_1 - x_3) \\ \frac{1}{2S}(x_2 - x_1) \end{array} \right) \left(\begin{array}{ccc} \mathcal{N}_1(x,y) & \mathcal{N}_2(x,y) & \mathcal{N}_3(x,y) \end{array} \right) dxdy$$

$$= \frac{1}{6} \left(\begin{array}{ccc} x_{32} & x_{32} & x_{32} \\ x_{13} & x_{13} & x_{13} \\ x_{21} & x_{21} & x_{21} \end{array} \right) \tag{C.72}$$

We now turn to the other two matrices, the advection $\mathbb{K}_a$ and diffusion $\mathbb{K}_d$ matrices. The gradient matrix $\boldsymbol{B}$ is given by

$$\boldsymbol{B} = \left(\begin{array}{ccc} \partial_x \mathcal{N}_1 & \partial_x \mathcal{N}_2 & \partial_x \mathcal{N}_3 \\ \partial_y \mathcal{N}_1 & \partial_y \mathcal{N}_2 & \partial_y \mathcal{N}_3 \end{array} \right) = \frac{1}{2S} \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right)$$

then

$$\boldsymbol{K}_d = \iint_{\triangle} \boldsymbol{B}^T k \boldsymbol{B} \, dV = \iint_{\triangle} \frac{k}{4S^2} \left(\begin{array}{cc} y_{23} & x_{32} \\ y_{31} & x_{13} \\ y_{12} & x_{21} \end{array} \right) \cdot \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right) dV$$

If k is constant within the element, then

$$\boldsymbol{K}_d = \frac{k}{4S} \left(\begin{array}{cc} y_{23} & x_{32} \\ y_{31} & x_{13} \\ y_{12} & x_{21} \end{array} \right) \cdot \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right)$$

Turning now to the advection matrix

$$\boldsymbol{K}_a = \iint_{\triangle} \vec{\mathcal{N}}^T \vec{v} \cdot \boldsymbol{B} \, dV$$

$$= \iint_{\triangle} \vec{\mathcal{N}}^T(x,y) \vec{v}(x,y) \cdot \boldsymbol{B}(x,y) \, dxdy$$

$$= 2S \iint_{\triangle} \vec{\mathcal{N}}^T(x(r,s), y(r,s)) \vec{v}(x(r,s), y(r,s)) \cdot \boldsymbol{B}(x(r,s), y(r,s)) \, drds$$

$$= 2S \iint_{\triangle} \left(\begin{array}{c} 1-r-s \\ r \\ s \end{array} \right) \vec{v}(x(r,s), y(r,s)) \cdot \frac{1}{2S} \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right) drds$$

$$= \iint_{\triangle} \left(\begin{array}{c} \mathcal{N}_1(r,s) \\ \mathcal{N}_2(r,s) \\ \mathcal{N}_3(r,s) \end{array} \right) \vec{v}(x(r,s), y(r,s)) \cdot \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right) drds$$

If the velocity is constant within the element (rather rare case) then this can be integrated exactly. If not, a quadrature rule must be used.

Let us assume that indeed velocity is constant inside the element. Then $\vec{v}(x(r,s), y(r,s)) = (u_0, v_0)$ and then

$$\boldsymbol{K}_a = \iint_{\triangle} \left(\begin{array}{c} \mathcal{N}_1(r,s) \\ \mathcal{N}_2(r,s) \\ \mathcal{N}_3(r,s) \end{array} \right) (u_0, v_0) \cdot \left(\begin{array}{ccc} y_{23} & y_{31} & y_{12} \\ x_{32} & x_{13} & x_{21} \end{array} \right) drds$$

$$= \iint_{\triangle} \left(\begin{array}{c} \mathcal{N}_1(r,s) \\ \mathcal{N}_2(r,s) \\ \mathcal{N}_3(r,s) \end{array} \right) \left(\begin{array}{ccc} u_0 y_{23} + v_0 x_{32} & u_0 y_{31} + v_0 x_{13} & u_0 y_{12} + v_0 x_{21} \end{array} \right) drds$$

$$\tag{C.73}$$

Using Eqs. (C.69),(C.70),(C.71) we arrive at

$$
\boldsymbol{K}_a = \frac{S}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} u_0 y_{23} + v_0 x_{32} & u_0 y_{31} + v_0 x_{13} & u_0 y_{12} + v_0 x_{21} \end{pmatrix} \tag{C.74}
$$

verify!! FOR REAL 2025 probably off by factor S/2 or 2S !!!

$$C_1 = \int_{\partial\Omega_1} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma$$

The edge $\partial\Omega_1$ is bounded by the coordinates of nodes x_1, y_1 and x_2, y_2. This segment can be parameterised by $t \in [0,1]$:

$$\vec{r}(t) = (1-t)\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + t\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} (x_2 - x_1)t + x_1 \\ (y_2 - y_1)t + y_1 \end{pmatrix}$$

Let us assume that C is a smooth curve and that it is given by the parametric equations $x = h(t)$, $y = g(t)$ and $a \le t \le b$. The line integral of a function $f(x,y)$ over C is computed as follows.

$$\int_C f(x,y)ds = \int_a^b f(h(t),g(t))\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\,dt$$

In our case $dx/dt = x_2 - x_1$ and $dy/dt = y_2 - y_1$ so

$$\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = L_1$$

Then

$$C_1 = \int_{\partial\Omega_1} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma$$

$$= \begin{pmatrix} \int_{\partial\Omega_1} N_1(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_1(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_1(x,y)N_3(x,y)d\Gamma \\ \int_{\partial\Omega_1} N_2(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_2(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_2(x,y)N_3(x,y)d\Gamma \\ \int_{\partial\Omega_1} N_3(x,y)N_1(x,y)d\Gamma & \int_{\partial\Omega_1} N_3(x,y)N_2(x,y)d\Gamma & \int_{\partial\Omega_1} N_3(x,y)N_3(x,y)d\Gamma \end{pmatrix}$$

$$= L_1 \begin{pmatrix} \int_0^1 N_1(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_1(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_1(x(t),y(t))N_3(x(t),y(\\ \int_0^1 N_2(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_2(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_2(x(t),y(t))N_3(x(t),y(\\ \int_0^1 N_3(x(t),y(t))N_1(x(t),y(t))dt & \int_0^1 N_3(x(t),y(t))N_2(x(t),y(t))dt & \int_0^1 N_3(x(t),y(t))N_3(x(t),y(\end{pmatrix}$$

We are about to compute the individual terms of the matrix one by one but we will need:

$$
\begin{aligned}
S &= \frac{1}{2}[(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)] \\
&= \frac{1}{2}[x_1 y_2 - x_1 y_3 - x_3 y_2 + x_3 y_3 - x_2 y_1 + x_2 y_3 + x_3 y_1 - x_3 y_3] \\
&= \frac{1}{2}[x_1 y_2 - x_1 y_3 - x_3 y_2 - x_2 y_1 + x_2 y_3 + x_3 y_1]
\end{aligned}
$$

$$(\text{C.76})$$

and

$$N_1(x(t), y(t)) = \frac{1}{2S}[x_2 y_3 - x_3 y_2 + (y_2 - y_3)x(t) + (x_3 - x_2)y(t)]$$

$$= \frac{1}{2S}[x_2 y_3 - x_3 y_2 + y_{23}(x_{21}t + x_1) + x_{32}(y_{21}t + y_1)]$$

$$= \frac{1}{2S}[x_2 y_3 - x_3 y_2 + y_{23}x_1 + x_{32}y_1 + (y_{23}x_{21} + x_{32}y_{21})t]$$

$$= \frac{1}{2S}[\underbrace{x_2 y_3 - x_3 y_2 + x_1 y_2 - x_1 y_3 + x_3 y_1 - x_2 y_1}_{2S} \tag{C.77}$$

$$+(x_2 y_2 - x_1 y_2 - x_2 y_3 + x_1 y_3 + x_3 y_2 - x_3 y_1 - x_2 y_2 + x_2 y_1)t]$$

$$= \frac{1}{2S}[2S - (\underbrace{x_1 y_2 + x_2 y_3 - x_1 y_3 - x_3 y_2 + x_3 y_1 - x_2 y_1}_{2S})t]$$

$$= \frac{1}{2S}[2S - 2St]$$

$$= 1 - t \tag{C.78}$$

$$N_2(x(t), y(t)) = \frac{1}{2S}[x_3 y_1 - x_1 y_3 + (y_3 - y_1)(x_{21}t + x_1) + (x_1 - x_3)(y_{21}t + y_1)]$$

$$= \frac{1}{2S}[x_3 y_1 - x_1 y_3 + (y_3 - y_1)x_1 + (x_1 - x_3)y_1 + (y_{31}x_{21} + x_{13}y_{21})t]$$

$$= \frac{1}{2S}[x_3 y_1 - x_1 y_3 + x_1 y_3 - x_1 y_1 + x_1 y_1 - x_3 y_1 + (y_{31}x_{21} + x_{13}y_{21})t]$$

$$= \frac{1}{2S}(y_{31}x_{21} + x_{13}y_{21})t$$

$$= t \tag{C.79}$$

$$N_3(x(t), y(t)) = \frac{1}{2S}(x_1 y_2 - x_2 y_1 + (y_1 - y_2)x(t) + (x_2 - x_1)y(t))$$

$$= \frac{1}{2S}(x_1 y_2 - x_2 y_1 + (y_1 - y_2)(x_{21}t + x_1) + (x_2 - x_1)(y_{21}t + y_1))$$

$$= \frac{1}{2S}(x_1 y_2 - x_2 y_1 + (y_1 - y_2)x_1 + (x_2 - x_1)y_1 + (y_{12}x_{21} + x_{21}y_{21})t)$$

$$= \frac{1}{2S}(x_1 y_2 - x_2 y_1 + x_1 y_1 - x_1 y_2 + x_2 y_1 - x_1 y_1 + (y_{12}x_{21} - x_{21}y_{12})t)$$

$$= 0 \tag{C.80}$$

then

$$\int_0^1 N_1(x(t), y(t))N_1(x(t), y(t))dt = \int_0^1 (1-t)^2 dt = 1/3 \tag{C.81}$$

$$\int_0^1 N_1(x(t), y(t))N_2(x(t), y(t))dt = \int_0^1 (1-t)t\, dt = 1/6 \tag{C.82}$$

$$\int_0^1 N_1(x(t), y(t))N_3(x(t), y(t))dt = 0 \tag{C.83}$$

$$\int_0^1 N_2(x(t), y(t))N_2(x(t), y(t))dt = \int_0^1 t^2 dt = 1/3 \tag{C.84}$$

$$\int_0^1 N_2(x(t), y(t))N_3(x(t), y(t))dt = 0 \tag{C.85}$$

$$\int_0^1 N_3(x(t), y(t))N_3(x(t), y(t))dt = 0 \tag{C.86}$$

and finally

$$\boldsymbol{C}_1 = \int_{\partial \Omega_1} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma = \frac{L_1}{6} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{C.87}$$

$$\boldsymbol{C}_2 = \int_{\partial \Omega_2} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma = \frac{L_2}{6} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \tag{C.88}$$

$$\boldsymbol{C}_3 = \int_{\partial \Omega_3} \vec{N}^T(x,y)\vec{N}(x,y)d\Gamma = \frac{L_3}{6} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \tag{C.89}$$

Appendix D

Finite element terminology in various languages

English	French	Dutch
Finite Element Method	Méthode des éléments finis	Eindige-elementenmet
Finite Difference Method	Méthode des différences finies	Eindige-differentiemet
Finite Volume Method	Méthode des volumes finis	
Matrix	Matrice	
Heat transport eq.	Equation de transport de la chaleur	Warmtetransport verg
Momentum conservation eq.	équation de conservation du moment	Wet van behoud van
Mass conservation / continuity eq.		Continuïteitsvergelijki
Iterative solver	solveur itératif	
Elemental matrix		
Boundary conditions	conditions aux limites	randvoorwaarden
(In)compressible	(in)compressible	
Surface processes	processus de surface	
an element	un élément	
Computational geodynamics	géodynamique numérique	
Assembly	assemblage	
Strong form		
Weak form	formulation variationnelle / formulation faible	
Basis function		
Shape function		
Partial differential eq. (PDE)	équation aux dérivées partielles (EDP)	partiële differentiaalve
Node	noeud	knooppunt
Grid, mesh	(la) maille / (le) maillage	rooster
Stiffness matrix	matrice de raideur	stijfheidsmatrix
Displacement vector	vecteur déplacement	verplaatsingsvector
Tessellation	pavage	Betegeling
Mass matrix	matrice de masse	
Classical mechanics	mécanique Newtonienne	(de) klassieke mechan
Momentum	(le) moment	(de) impuls
Perimeter	(le) perimetre	(de) omtrek
Wavelength	(la) longueur d'onde	de golflengte

Appendix E

Fun modelling

Because sometimes numerical modelling *is* fun ...
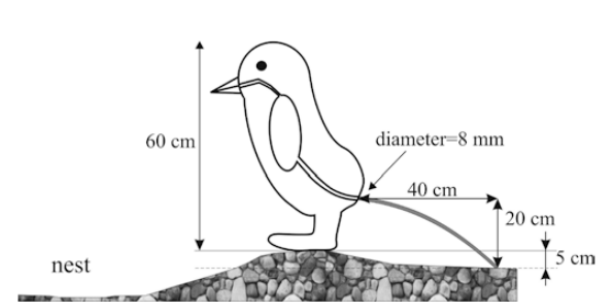
Pressures produced when penguins pooh - calculations on avian

defaecation [918]

Clothes washing simulations [3]

DON'T SUPPRESS THE WIGGLES—THEY'RE TELLING YOU SOMETHING!†

PHILIP M. GRESHO and ROBERT L. LEE

[515]

ARE FEM SOLUTIONS OF INCOMPRESSIBLE FLOWS REALLY INCOMPRESSIBLE? (OR HOW SIMPLE FLOWS CAN CAUSE HEADACHES!)

DOMINIQUE PELLETIER, ANDRE FORTIN AND RICARDO CAMARERO

[1045]
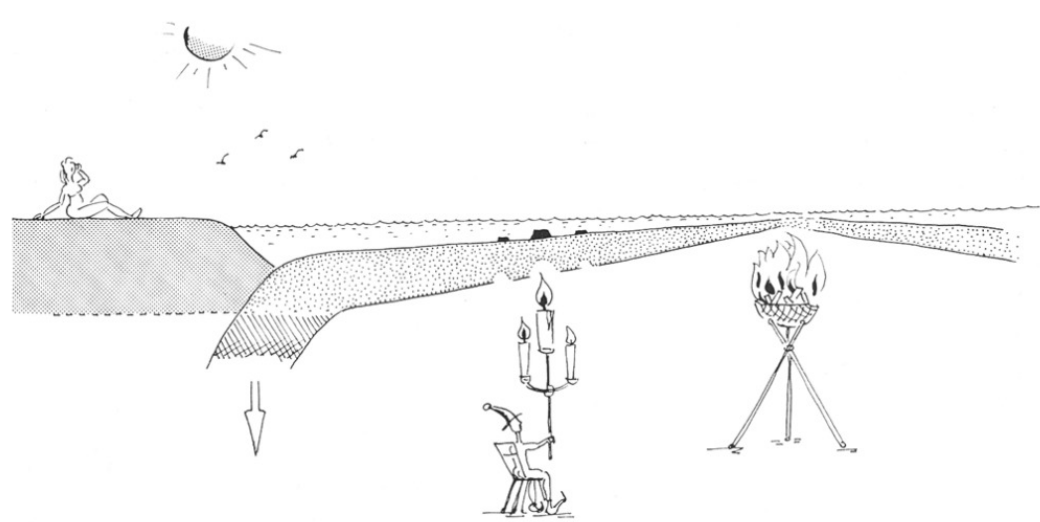
Lithospheric thickness anomaly near the trench and possible driving force of subduction [448]
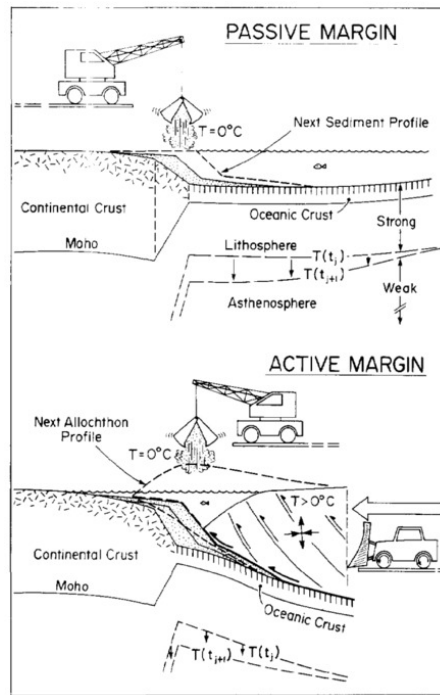
Convergent margin tectonics [1299]

INCOMPRESSIBILITY WITHOUT TEARS—HOW TO AVOID RESTRICTIONS OF MIXED FORMULATION

O. C. ZIENKIEWICZ AND J. WU

Institute of Numerical Methods in Engineering, University College of Swansea, Swansea, U.K.

[1545]

Collective Motion of Humans in Mosh and Circle Pits at Heavy Metal Concerts

Jesse L. Silverberg,* Matthew Bierbaum, James P. Sethna, and Itai Cohen
Department of Physics and Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA
(Received 13 February 2013; published 29 May 2013)

Human collective behavior can vary from calm to panicked depending on social context. Using videos publicly available online, we study the highly energized collective motion of attendees at heavy metal concerts. We find these extreme social gatherings generate similarly extreme behaviors: a disordered gaslike state called a *mosh pit* and an ordered vortexlike state called a *circle pit*. Both phenomena are reproduced in flocking simulations demonstrating that human collective behavior is consistent with the predictions of simplified models.

[1244]

Toasting the jelly sandwich: The effect of shear heating on lithospheric geotherms and strength

Ebbe H. Hartz   Physics of Geological Processes, University of Oslo, 0316 Oslo, Norway, and Aker Exploration,
Haakon VII's gt. 9, P.O. Box 580, Sentrum, NO, 4003 Stavanger, Norway
Yuri Y. Podladchikov   Physics of Geological Processes, University of Oslo, 0316 Oslo, Norway

[584]
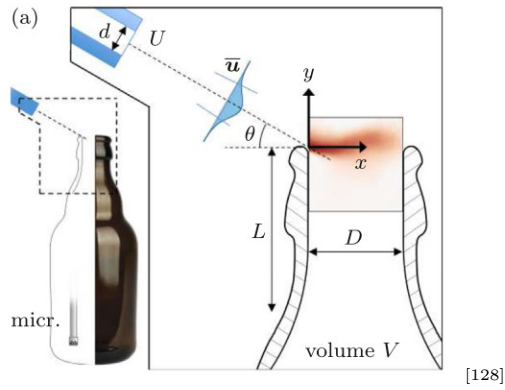
[128]

Effect of soccer shoe upper on ball behaviour in curve kicks [657]

[1420]

[1120]

[1205]

1289

Appendix F

Beautiful/interesting images from computational geodynamics

[1529]

[1272]

[322]

[440]

[916]

[482]

flat slabs absent

flat slabs dominate
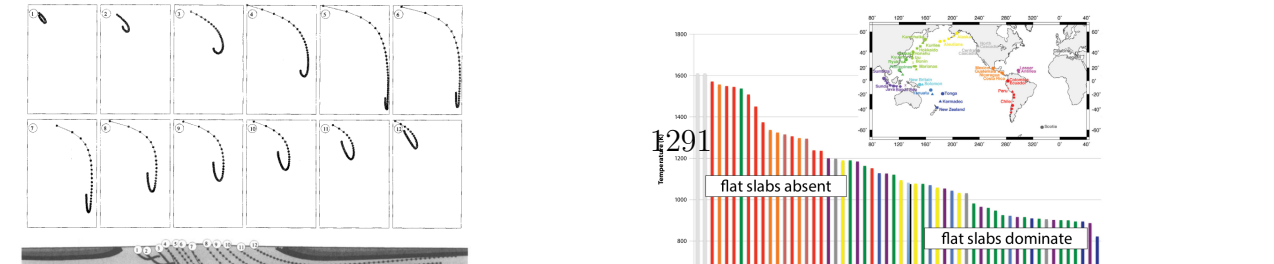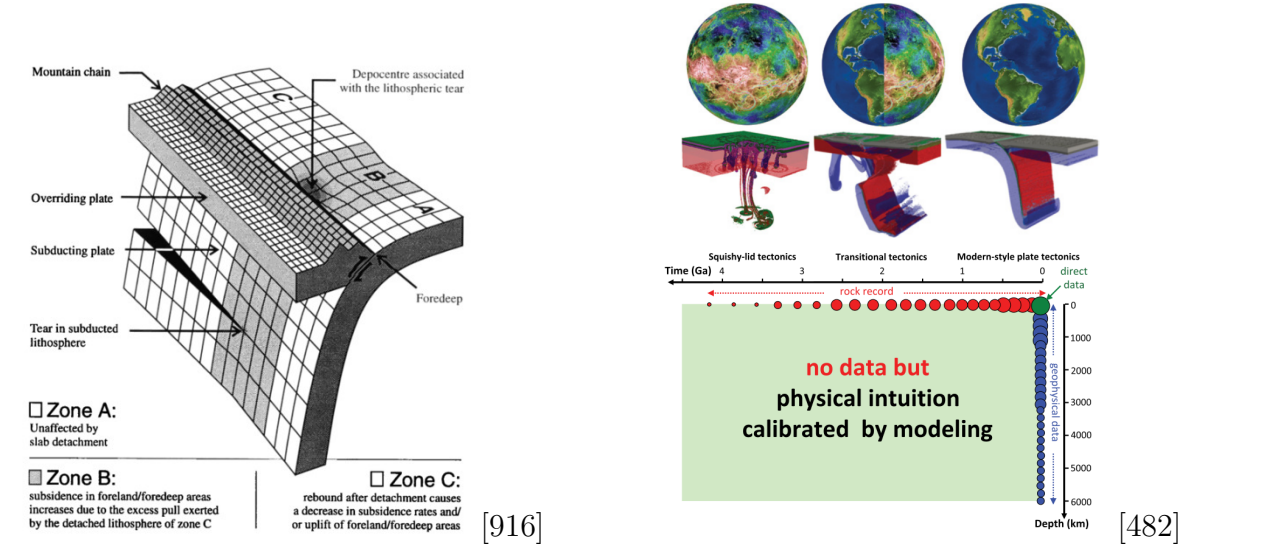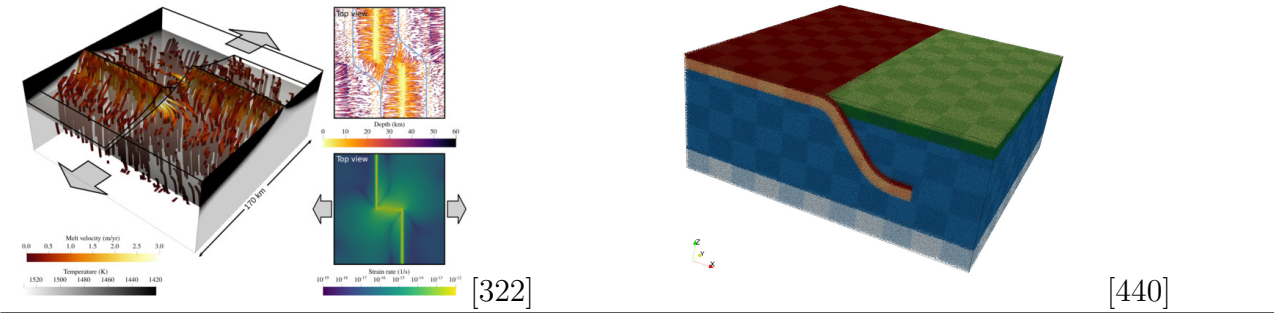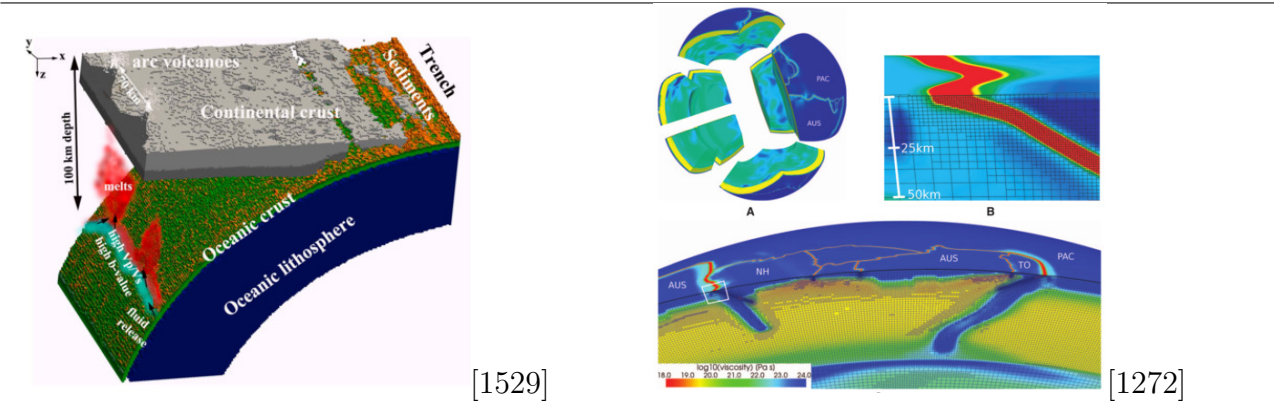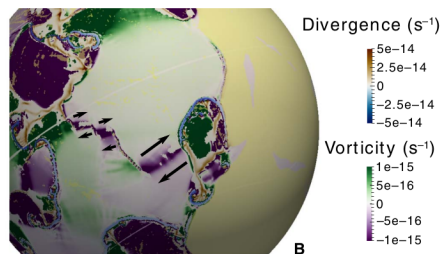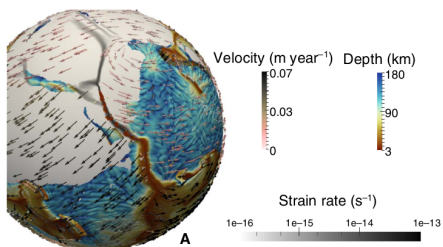
[291]
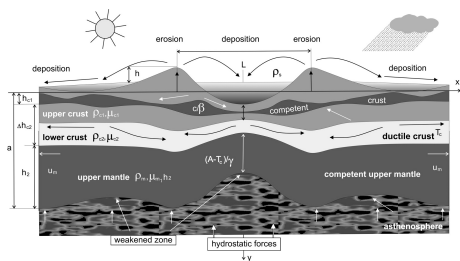
[197]

[1051]

[310]

[1324]

[1242]

[1005]

1292

b

Lithosphere

Serpentine

— 100

UPPER MANTLE

Oceanic Crust

Olivine

— 410

Earthquakes stop here

Spinel

$\Delta V = -6\%$

MANTLE TRANSITION ZONE

— 660 km

Perovskite + oxide $\Delta V = -8\%$

LOWER MANTLE

[513]

continental plate

oceanic plate

suction force (active back-arc spreading)

normal force at transform fault

tangential force at transform fault

ridge push

crustal potential energy differences

colliding resistance

viscous drag

lithospheric thickening

viscous drag

mantle density driven convection

slab pull

subduction resistance

distributed forces based on flow calculations

edge forces based on orientation of plate boundary and relative velocities

distributed lithospheric forces based on isostatic potential energy differences

[72]

PLATE

[618]

TRENCH

VOLCANIC LINE

HIGH HEAT FLOW, ANOMALOUS MANTLE

400°C

400°C

800°C

CONTINENTAL

100 km

800°C

1200°C

OCEANIC

1200°C

300 km

1600°C

OLIVINE SPINEL

1600°C

500 km

1700°C

SPINEL OXIDES

1700°C

700 km

900 km

[1220]

1293

Appendix G

Working with Git from the terminal

G.0.1 Contributing to Aspect

1. make sure that you have an account on GitHub (say, `https://github.com/myusername`) and carry out a proper setup on your local computer as follows:
   `$ git config --global user.name "firstname lastname"`
   `$ git config --global user.email your@email`

2. On github.com, fork the official ASPECT repository to your repository:
   go to `https://github.com/geodynamics/aspect` and click on the 'fork' button on the upper right corner of the screen.

3. On your machine, in a terminal, clone your repository with
   `$ git clone git@github.com:myusername/aspect.git`. This is now your main[1] branch.

4. Follow the instructions at `https://docs.github.com/en/authentication/connecting-to-github-w` `adding-a-new-ssh-key-to-your-github-account` to add a new SSH key to your GitHub account.

5. create a remote of your (online) repository
   `$ git remote add origin git@github.com:myusername/aspect.git` and in order to avoid potential confusion later on, we shall rename our github repo as follows:
   `$ git remote rename origin myusername`

6. Also create a remote of the official version with
   `$ git remote add upstream https://github.com/myusername/aspect`

7. Do `$git remote -v` which shows you the URLs that git has stored for the shortname to be used when reading and writing to that remote.

   ```
   $ git remote -v
   myusername git@github.com:myusername/aspect.git (fetch)
   myusername git@github.com:myusername/aspect.git (push)
   upstream https://github.com/geodynamics/aspect.git (fetch)
   upstream https://github.com/geodynamics/aspect.git (push)
   ```

---

[1]previously 'master', although this term should not be used anymore
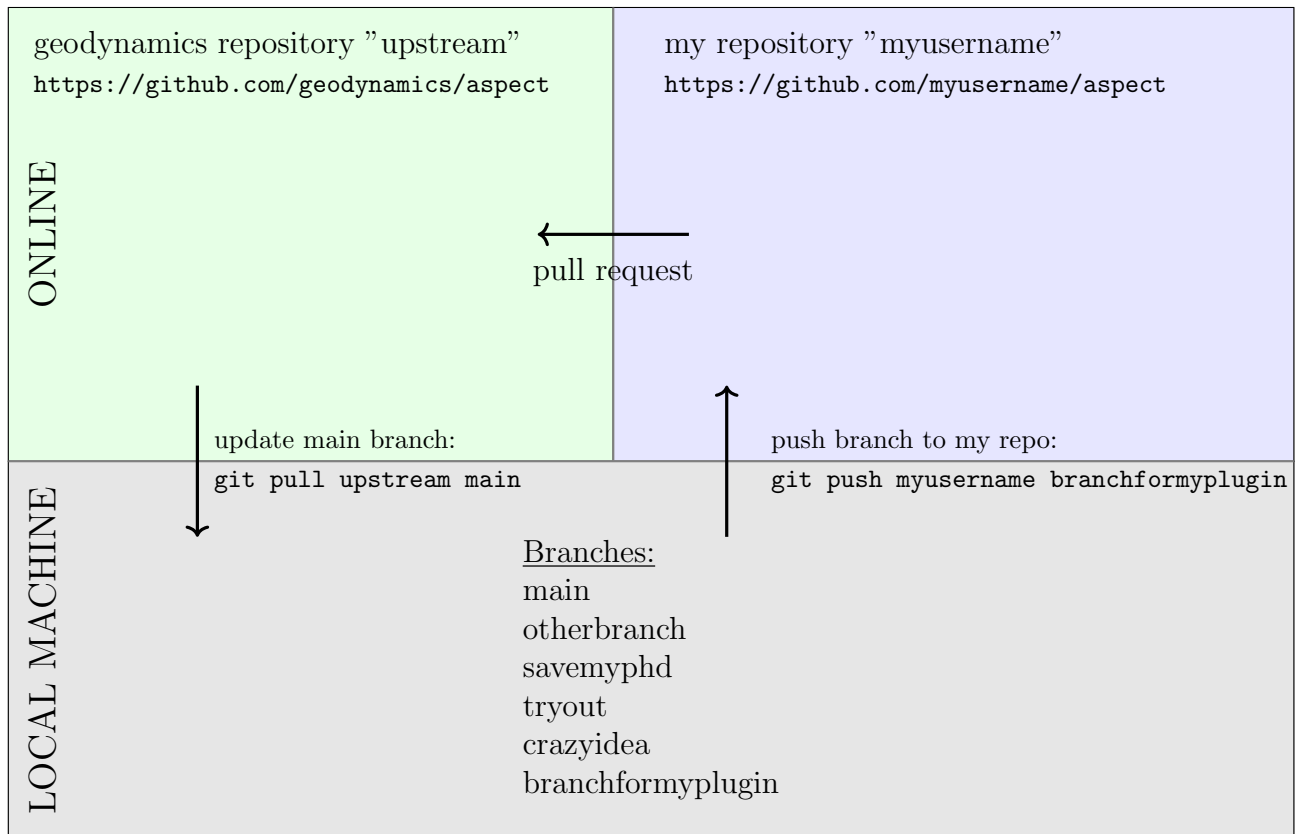
Very concretely, if you wish to contribute: Let us assume that you have found a typo in the file "physics.cc" and you wish to fix this problem.

1. If this is the very first time you use git, the following instruction is not needed. If not, make sure that your terminal points at your main branch: `$ git checkout main`

2. Then, make sure you update your local version: `$ git pull upstream main` and also update your own repo online: `$ git push username main`

3. Create a new branch with a self explanatory name: `$ git branch fix_typo` Then do `$ git branch` to see all your branches. The one that is coloured is the active branch. In order to switch to this new branch, do `$ git checkout fix_typo`. Redo `$ git branch` to verify that the branch 'fix_typo' is highlighted.

4. Edit the file physics.cc, correct the typo, save and exit. Do `$ git status` and you should see the filename highlighted next to 'modified'.

5. before we go any further, we need to run the indenting script with astyle. In the build directory run `$ make indent`. Note that you need a specific version of astyle, see `https://github.com/geodynamics/aspect/wiki/Indenting-code-and-installing-the-correct-version-of-astyle`

6. add a changelog entry in doc/modules/changes. Look at the entries and find the one that resembles your contribution the most. Use it to write your entry.

7. If this is the only modification you wish to communicate, you then need to add and commit it as follows:

   `$ git add physics.cc`

   If you do `$ git status` again, the file should have changed colour. Then do

1295

```
$ git commit -m "message"
```

where 'message' is a very short description of the modification (e.g. 'I fixed a typo').

8. We now need to bring this modification online

```
$ git push myusername fix_typo
```

9. You are nearly done. The last step takes place on github.com: when you log onto your own github fieldstone page, you should see a large green rectangle "Compare and pull request". Click on this button, follow the instructions.

10. a new page opens, entitled "Open a pull request" (PR). If necessary, add a detailed description of the pull request (this makes sense when you contribute a piece of code, or a whole new section, etc ...). Later you will be able to review the requested changes at the bottom. In the end, simply click "Create pull request".

11. Once you have done so, Aspect developers will then review it. If they have no comment, the PR will be accepted and your modification will then be incorporated in the main repo. If they have comments, you will be notified via github and a back-and-forth discussion will ensue until the PR is accepted.

12. if the reviewers have comments. Edit/change your file(s). Then add these (`git add ...`), and commit again (`git commit -m "msg"`) and push as before.
OR: better:
`git commit --amend -a` so that you don't have to squash/fixup afterwards

13. After the PR has been accepted, the branch is no longer needed. Switch back to your local main branch: $ `git checkout main` . Update your local main and online repo (see step # 2) and then delete the no-longer-needed branch as follows:

```
$ git branch -d fix_typo
```

———————————————

If there is a pb with your PR and u need to rebase. For example if your 2 PRs modify the same line (say for example reference.bib - in that case better spread your changes to different locations in the reference.bib)

- carry out modifications as required by reviewers

- `git add` files, and `git commit`

- `git checkout main`

- `git branch`, make sure you are indeed back on main

- `git pull upstream main`. depending how much happened in the last hours/days it will display a bunch of files/updates

- `git checkout my_branch`

- `git rebase main`. Follow instructions, resolve conflicts in indicated files. `git add problematic_file`. Finish with `git rebase --continue`.

- `git push -f cedrict my_branch`

To configure the editor used by git (do it once):

`git config --global core.editor "vim"`

git stash git stash pop
git log

how to squash/fixup:

git rebase -i main − >

A similar window as this will open:

Since we wish to squash (or rather fixup!) we do:

save and exit

do git diff

G.0.2 Contributing a cookbook in Aspect

1. Download/update/install ASPECT

2. In the cookbooks folder, create a new folder: `$ mkdir my_cool_setup`

3. In this folder, place your .prm file, say my_cool_setup.prm

4. Make sure your .prm file is clean, commented, and contains a header with a concise description of what the experiment is, and/or in which publication it originates.

5. In the folder, create a new folder: `$ mkdir doc`

6. In this folder create the file my_cool_setup.md file which contains the text for the cookbook. Look at other cookbooks md files for examples of how to include a figure, an equation, cite publications, etc ...

7. Place in this same folder all figures pertaining to the cookbook entry in the manual.

8. go to `/doc/sphinx/user/cookbooks/` and add your cookbook to the list in (for example) `geophysical-setups.md`

9. if you with to cite publications, add them to `/doc/sphinx/references.bib`

10. In order to generate the manual, go to /doc/sphinx and do `$ make html`

11. if you wish to re-generate the part of the manual that comes from the documentation of .cc files, then got to `\build` and `make` the code, then do in `/doc/`:

    `./update_parameters.sh /home/absolute/path/aspect/build/aspect`

    and then do `make html`. If you re-modify the .cc file, you need to redo all 3 steps.

12. If there is no error, you should be able to open the file `/doc/sphinx/\_build/html/index.html` with firefox

13. if the referencing of the figures does not work correctly, simply do `make clean` and then make html again.

14. before you make a pull request, make sure you run `make indent`

G.0.3   Contributing to fieldstone

This appendix was contributed by E. van der Wiel.

1. make sure that you have an account on GitHub (say, `https://github.com/myusername`) and carry out a proper setup on your local computer as follows:
   `$ git config --global user.name "firstname lastname"`
   `$ git config --global user.email your@email`

2. On github.com, fork the official fieldstone repository to your repository:
   go to `https://github.com/cedrict/fieldstone` and click on the 'fork' button on the upper right corner of the screen.

3. On your machine, in a terminal, clone your repository with
   `$ git clone git@github.com:myusername/fieldstone.git`. This is now your main[2] branch.

4. On your machine, find your security key [3] with `$ less ~/.ssh/id_dsa.pub` and copy this into github.com so you can push to your repository. See `https://help.github.com/en/articles/connecting-to-github-with-ssh` on how to configure github with ssh support (no more login/password to type – if you have cloned the repository wish ssh, not html). Please also check the instructions at `https://help.github.com/en/articles/connecting-to-github-with-ssh`.

5. create a remote of your (online) repository
   `git remote add origin git@github.com:myusername/fieldstone.git` and in order to avoid potential confusion later on, we shall rename our github repo as follows:
   `$ git remote rename origin myusername`

6. Also create a remote of the official version with
   `$ git remote add upstream https://github.com/cedrict/fieldstone`

---

[2]also 'master', although this term should not be used anymore

[3]Note that you can create a public key as follows: `https://help.github.com/articles/generating-ssh-keys/`
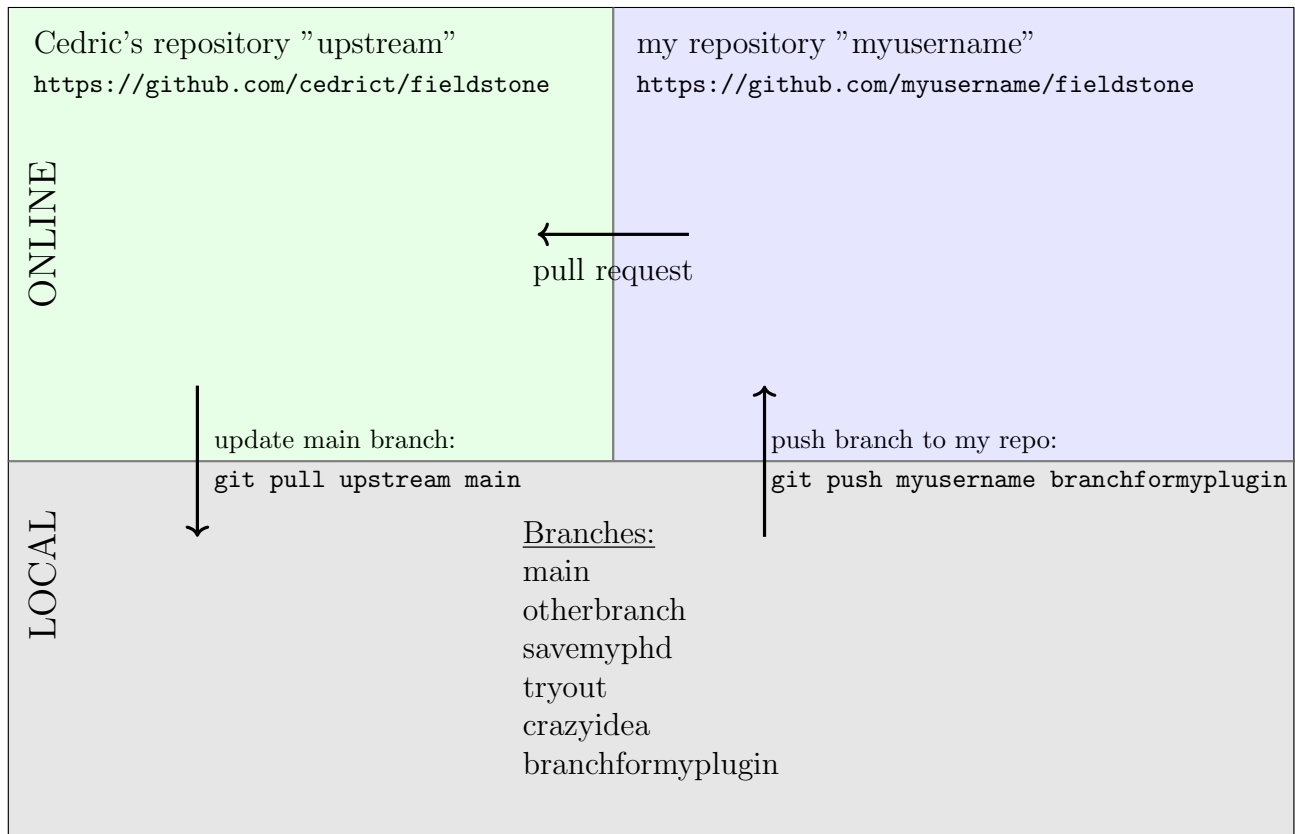
| Cedric's repository "upstream" | my repository "myusername" |
| https://github.com/cedrict/fieldstone | https://github.com/myusername/fieldstone |

ONLINE

pull request

update main branch:
`git pull upstream main`

push branch to my repo:
`git push myusername branchformyplugin`

LOCAL

Branches:
main
otherbranch
savemyphd
tryout
crazyidea
branchformyplugin

Very concretely, if you wish to contribute: Let us assume that you have found a typo in the file "physics.tex" and you wish to report the problem to me. The easiest way is to write me an email with the nature of the problem and the proposed fix. Although I will be grateful for your contribution, this approach can be improved upon by using git's "pull request" command.

1. If this is the very first time you use git, the following instruction is not needed. If not, make sure that your terminal points at your main branch: `$ git checkout master`

2. Then, make sure you update your local version: `$ git pull upstream master` and also update your own repo online: `$ git push`

3. Create a new branch with a self explanatory name: `$ git branch fix_typo` Then do `$ git branch` to see all your branches. The one that is coloured is the active branch. In order to switch to this new branch, do `$ git checkout fix_typo`. Redo `$ git branch` to verify that the branch 'fix_typo' is highlighted.

4. Edit the file and correct the typo, save and exit. Do `$ git status` and you should see the filename highlighted next to 'modified'.

5. If this is the only modification you wish to communicate, you then need to add and commit it as follows:

   `$ git add physics.tex`

   If you do git status again, the file should have changed colour (?). Then do

   `$ git commit -m "message"`

   where 'message' is a very short description of the modification (e.g. 'I fixed a typo').

6. We now need to bring this modification online

   ```
   $ git push myusername fix_typo
   ```

7. You are nearly done. The last step takes place on github.com: when you log onto your own github fieldstone page, you should see a large green rectangle "Compare and pull request". Click on this button.

8. a new page opens, entitled "Open a pull request". If necessary, add a detailed description of the pull request (this makes sense when you contribute a piece of code, or a whole new section, etc ...). You can review the requested changes at the bottom. In the end, simply click "Create pull request".

9. Once you have done so, I will receive an email which notifies me of the pull request. I will then review it. If I have no comment, I will accept the PR and your modification will then be incorporated in the master repo. If I have comments, you will be notified via github and a back-and-forth discussion will ensue until I accept the PR.

10. After the PR has been accepted, the branch is no longer needed. Switch back to your local master branch: $ `git checkout master` . Update your local master and online repo (see step # 2) and then delete the no-longer-needed branch as follows:

    ```
    $ git branch -d fix_typo
    ```

```
(base) UU062931:fieldstone vanderWiel$ git branch
  master
* my_work
(base) UU062931:fieldstone vanderWiel$ git checkout master
Switched to branch 'master'
Your branch is up to date with 'origin/master'.
(base) UU062931:fieldstone vanderWiel$ git pull upstream master
From https://github.com/cedrict/fieldstone
 * branch            master     -> FETCH_HEAD
Already up to date.
(base) UU062931:fieldstone vanderWiel$ git push
Everything up-to-date
(base) UU062931:fieldstone vanderWiel$
```

```
(base) UU062931:fieldstone vanderWiel$ git branch fix_typo
(base) UU062931:fieldstone vanderWiel$ git branch
  fix_typo
* master
  my_work
(base) UU062931:fieldstone vanderWiel$ git checkout fix_typo
Switched to branch 'fix_typo'
(base) UU062931:fieldstone vanderWiel$ git branch
* fix_typo
  master
  my_work
(base) UU062931:fieldstone vanderWiel$
```

```
(base) UU062931:fieldstone vanderWiel$ vi physics.tex
(base) UU062931:fieldstone vanderWiel$ git status
On branch fix_typo
Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
        modified:   physics.tex

no changes added to commit (use "git add" and/or "git commit -a")
(base) UU062931:fieldstone vanderWiel$ git add physics.tex
(base) UU062931:fieldstone vanderWiel$ git status
On branch fix_typo
Changes to be committed:
  (use "git restore --staged <file>..." to unstage)
        modified:   physics.tex

(base) UU062931:fieldstone vanderWiel$ git commit -m "I fixed a typo"
[fix_typo 85beb3e] I fixed a typo
 1 file changed, 1 insertion(+), 1 deletion(-)
(base) UU062931:fieldstone vanderWiel$ git push origin fix_typo
Enumerating objects: 5, done.
Counting objects: 100% (5/5), done.
Delta compression using up to 4 threads
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 294 bytes | 294.00 KiB/s, done.
Total 3 (delta 2), reused 0 (delta 0)
remote: Resolving deltas: 100% (2/2), completed with 2 local objects.
remote:
remote: Create a pull request for 'fix_typo' on GitHub by visiting:
remote:      https://github.com/e-wiel/fieldstone/pull/new/fix_typo
remote:
To https://github.com/e-wiel/fieldstone.git
 * [new branch]      fix_typo -> fix_typo
(base) UU062931:fieldstone vanderWiel$
```

| ⓟ 235 commits | ⑂ 3 branches | ♡ 0 releases | ⚑ 3 contributors | ⚖ View license |

Your recently pushed branches:

⑂ fix_typo (1 minute ago)   ⑂ Compare & pull request

| Branch: master ▼ | New pull request | | Create new file | Upload files | Find file | Clone or download ▼ |

This branch is even with cedrict:master.   ⑂ Pull request ⓘ Compare

| 🖼 cedrict updatekjh | | Latest commit e6faecc 41 minutes ago |

| ▣ images | bibl243 | 3 hours ago |

```
(base) UU062931:fieldstone vanderWiel$ git branch
  fix_typo
* master
  my_work
(base) UU062931:fieldstone vanderWiel$ git pull upstream master
remote: Enumerating objects: 6, done.
remote: Counting objects: 100% (6/6), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 6 (delta 3), reused 4 (delta 3), pack-reused 0
Unpacking objects: 100% (6/6), done.
From https://github.com/cedrict/fieldstone
 * branch            master     -> FETCH_HEAD
   e6faecc..cb3a879  master     -> upstream/master
Updating e6faecc..cb3a879
Fast-forward
 github.tex  | 147 +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 manual.pdf  | Bin 65759829 -> 65769253 bytes
 physics.tex |   2 +-
 3 files changed, 143 insertions(+), 6 deletions(-)
(base) UU062931:fieldstone vanderWiel$ git branch -d fix_typo
Deleted branch fix_typo (was 85beb3e).
(base) UU062931:fieldstone vanderWiel$ git push
Total 0 (delta 0), reused 0 (delta 0)
To https://github.com/e-wiel/fieldstone.git
   e6faecc..cb3a879  master -> master
(base) UU062931:fieldstone vanderWiel$
```

Screen captures of the procedure described above as carried out by E. van der Wiel on his Apple laptop.

In what follows we summarize the most important commands one should and remember while working with github. After creating an account one can 'fork' a repository (repo) in the online environment. This repository is a copy from the master directory of the developer and should not be used to adapt or change, as changes from the developer (updates) should be obtained in this 'fork', or as it could also be called; your master branch.

In order to be able to work within a repository, for instance, to run and compile different programs, you should have you own branch of the repository in which YOU CAN make changes. The following commands should be used to make, copy and publish your own version of the repo to your local device and the online github environment.

command	what it does
git branch	shows all branches of your repository and highlights the one you're in.
git checkout -b <my_own_branch>	This makes your own branch called "my_own_branch".
git push origin <my_own_branch>	This pushes your own, local, branch to as a second branch in the online repo of github.
git checkout <name>	changing the branch your working in (e.g. master or my_own_branch). Or replace the name with a hyphen to switch to the last branch.
git branch -d <my_own_branch>	Delete your local branch.

The following commands should be used in order to update your own local branches from updates made by somewhere else (upstream/master is the main repository). One should do this for the local master branch and, where possible as well for the different local branches you have committed changes to already.

command	what it does
git checkout master	To make sure you are in the right branch
git fetch upstream	to fetch updates from upstream repositories to you own local branch (e.g. to update your master branch.
git merge upstream/master	Command to update the branch with the fethched repo from 'upstream'.
git push origin master	To level your own online repository again with the one on your local drive (and thus the one upstream).
git checkout <my_own_branch>	To switch to your own adapted branch of the repo.
git merge master	Used from another branch working directory to combine the new released version of the master repo with the one where all your own changes are put. ->Then git finds all conflicts in different files which you need to resolve.
git add .	This adds the resolved issues in your own local branch (not master). After which you are able to commit and push your changes back to the online respository.

While you are working in your own branch you can change, add or delete files in any amount you want. However, always check whether your changes do not inflict the outcome of for instance your code. And when uploading from your terminal: if you commit and then push from master branch

1304

your changes will automatically be inserted in the online version of your master branch, when done from another branch it will be shown as a pull request towards your master branch. This request can than, for instance be forwarded to the main repo.

command	what it does
git commit -a	This will send your changes/updates from your branch as a commit to your own local branch.
git push origin <changes>	To update the remote repository (on Github) from you local repository (in this case the 'changes' branch). (Actually upload the new version). Online one can then judge what to do with it. !! this is a pull request towards your own fork/local_branch
git status	Showing the status of your current branch; it shows which files are different between the master file and your adapted branch.
git diff <changes>	This shows the exact differences between the different branches; one can simply ask for the difference between two branches when pwd in one branch ask for the other branch.
git merge <my_own_branch>	When used from the master branch (or any other???) this accepts the changes made in your branch and puts them in your local(!) master branch.
git pull origin master	if the main repository changes, one can pull the newest version towards it's own master file. While keeping your own branches alive with you own changes and vica versa: by running this command the origin/master (remote file) will be cloned and updated to the working branch you are in.
git stash (apply)	?? While updating your local branch, sometimes git wants to overrule your own changes, with this command you can 'stash' them to look at the differences later. ??

Appendix H

Writing a report as homework

- The document should contain your full name and student number on the first page.

- The file should be a pdf which name contains your family name

- Layout: is the document visually pleasing? Is it well structured?

- Is there a complete bibliography (when applicable)?

- Does the structure follows this: Introduction - Methods - Results - Discussion - Conclusion - Appendix ?

- Figures: Are they properly numbered? captioned? all figures must be referenced in the text. Are they of good enough quality (no visible pixels)? are they readable? are all axis labelled?

- Text: Overall quality of the language. Are there still typos? Do all sentence make sense?

- if you wish to show lines of code, use verbatim or lstlisting[1]

- Discussion: are the results properly discussed, analyzed? are potential problems, errors, limitations discussed?

- Conclusion: Are the findings/results summarized and generalized?

No	Yes
$6.67 * 10^{-11}$	6.67×10^{-11}
kg/m^3	kg/m^3 or kg.m^{-3}
1x1	1$\times$1
cos	cos
docx file	pdf file
'if you do this'	passive form
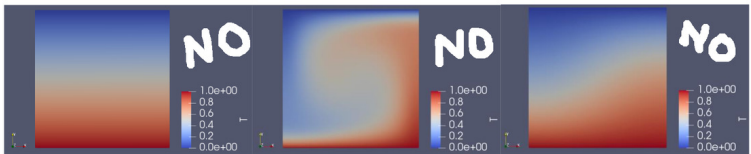
No grey background

---

[1] https://en.wikibooks.org/wiki/LaTeX/Source_Code_Listings

U = [-41346074.59639838, -41346106.14861266, -41346322.12423281, -41347034.8219668,
-41348728.85336469, -41352079.33063214, -41357972.05539607, -41367516.98181634
-41382034.29417382, -41402966.8138888, -41431618.19900567, -41468508.42097412
-41511962.23285898, -41555426.6099448, -41583595.20822245, -41570144.20903767
-41484197.88710789, -41308649.41597945, -41054103.1845507, -40748933.68245689
-40415944.45883375, -40057457.53015289, -39655212.74099992, -39185002.81130297
-38639470.67788275, -38036907.55824716, -37405072.1363776, -36759267.24928432
-36092497.74498677, -35380493.83127169, -34600914.0153607, -33754968.22552883
-32868819.25937854, -31975267.12561081, -31098665.54160668, -30252235.42279971
-29441444.64112505, -28667591.04175556, -27930018.48500311, -27227239.55989363
-26557456.02105067, -25918791.91839744, -25309397.59712495, -24727496.32338516
-24171404.76853348, -23639540.9187365, -23130425.50643475, -22642679.89858946
-22175021.99222236, -21726261.0159415, -21295291.79559261, -20881088.84329066
-20482700.50219241, -20099243.29398487, -19729896.55730399, -19373897.42481487
-19030536.15912975, -18699151.84940752, -18379128.45870995, -18069891.20501603
-17770903.2548138, -17481662.7063611, -17201699.83931143, -16930574.60791429
-16667874.35607352, -16413211.73392739, -16166222.7971429, -15926565.27168286
-15693916.96834184, -15467974.33280896, -15248451.11838825, -15035077.16976939
-14827597.30739603, -14625770.30303001, -14429367.93805357, -14238174.13690581
-14051984.16881316, -13870603.91165919, -13693849.17245316, -13521545.05940309
-13353525.40109213, -13189632.20869422, -13029715.17755753, -12873631.22483632
-12721244.06016446, -12572423.78664866, -12427046.52970981, -12284994.09152793

No lists/arrays with numbers
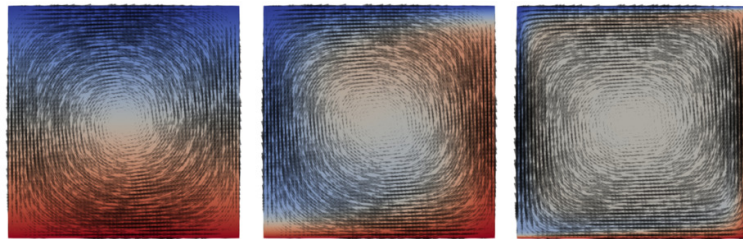
Too many arrows
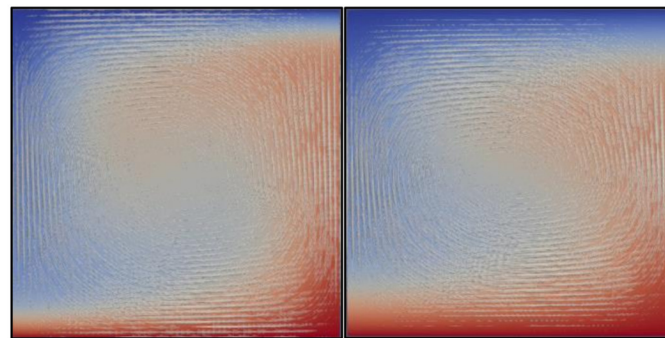
Poor choice of arrow colour
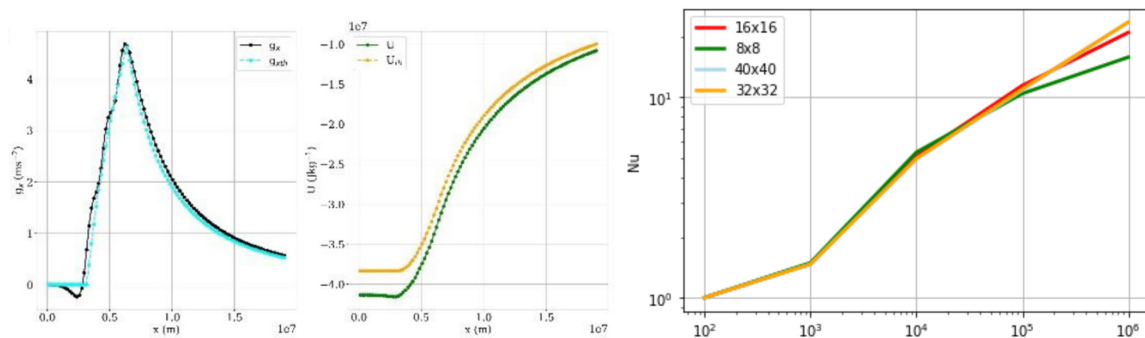
Be careful about how you export your figures. These are unreadable.

$$vrms = \left(\frac{1}{V_\Omega} \int_\Omega |\boldsymbol{v}|^2 \, dV\right)^{\frac{1}{2}}$$

$$= \left(\frac{1}{V_\Omega} \int_\Omega (u^2 + v^2) \, dV\right)^{\frac{1}{2}} \quad (9)$$

Parenthesis too small

Figure 1, steady state for the purpose of this work is to be considered within the interval corresponding to Ra values between 1E+04 and 1E+05. Hence, Nu values for a steady state range from 4.92E+00 to 1.14E+01 among all tested resolutions. For the model's standard 32x32 resolution, the Nu value rises from 4.94E+00 to 1.10E+01 within the Ra interval of 1E+04 to 1E+05. The critical Ra number (Ra$_c$) for all four resolutions seems to coincide at the value of 5E+02.

1.6E+10 is not acceptable. Replace by $1.6 \cdot 10^{10}$

$$\boldsymbol{\nabla} \cdot [\eta(\boldsymbol{\nabla v} + \boldsymbol{\nabla v}^T)] - \boldsymbol{\nabla} p + \rho(T)\boldsymbol{g} = \boldsymbol{0} \quad (5)$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0 \quad (6)$$

$$\rho_0 C_p \boldsymbol{v} \cdot \boldsymbol{\nabla} T = k\Delta T \quad (7)$$

Equation number is too close to the equation itself. Use labels, do not number equations by hand.
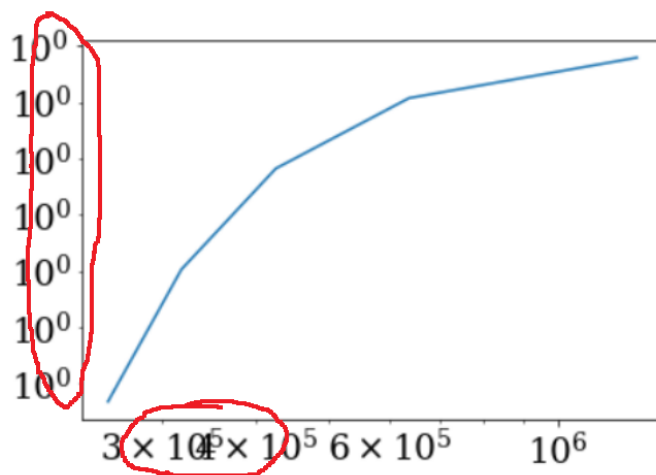
Formatting of both axis lead to unreadable figure.

$$g_x(x, y, z) = \mathcal{G}\Sigma_{e=1}^{N_e} \rho_e V_e \frac{x - x_e}{|\vec{r} - \vec{r_e}|^3}$$

In LaTeX use \sum\limits

The computed mass, using the PREM density distribution, the mass of the Earth, and the discrepancy between the two are 5.967863262629439e+24 kg,

Are so many digits necessary?

1308

With this new density a new mass for the sphere can be calculated. The mass calculated with the PREM density is $5.138471643448518*10^{24}kg$, the mass of the Earth is $5.972*10^{24}kg$, the difference between the calculates mass of the sphere and the mass of the Earth is $8.335283565514822*10^{23}kg$. From the mass calculated with PREM density the gravitational acceleration at the surface can also be calculated and compared between the sphere and the Earth. $g_{sphere} = 8.38566314096455m/s^2$, $g_{Earth} = 9.78291192747942m/s^2$, $g_{difference} = 1.39724878651487m/s^2$.

use \usepackage[cm]{fullpage} to allow for wider text.
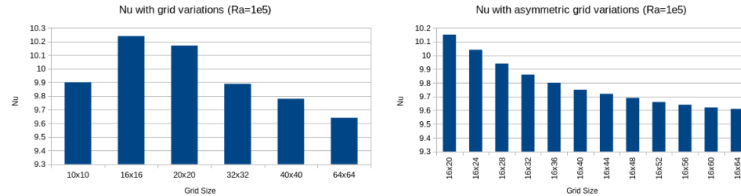
This figure style is to be avoided. Simply use dots and/or lines.
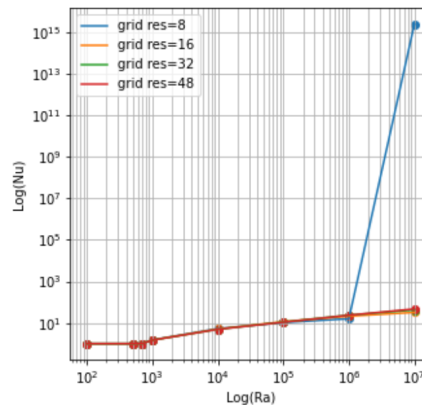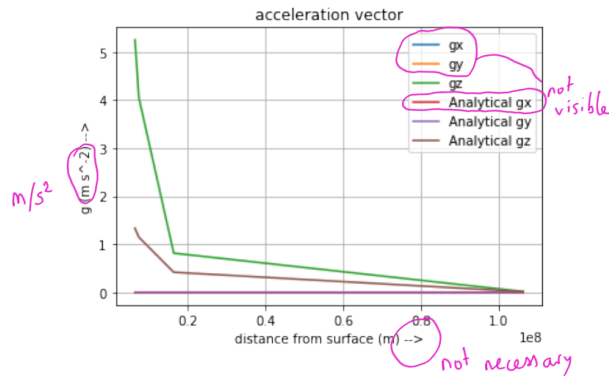
Here Ra and Nu are plotted in log-log scale, not log(Ra) and log(Nu).

The analytical solution for the gravitational field of a **solid sphere** with radius R is as follows:

$$g(r)_{int} = -\frac{4\pi}{3}G\rho_0 r; \tag{1}$$

$$U(r)_{int} = \frac{2\pi}{3}G\rho_0 r^2 + \frac{3}{2}\frac{GM}{r} \tag{2}$$

... for the spere's interior $(r \leq R)$, and ...

$$g(r)_{ext} = \frac{MG}{r^2}; \tag{3}$$

$$U(r)_{ext} = -\frac{GM}{r} \tag{4}$$

... for the sphere's exterior $(r \geq R)$, where $g(r)$ is the gravitional acceleration vector norm and $U(r)$ the gravita-

The dots at the beginning and end of the lines are not necessary.

Never *never* use 'you'.

For both the full and hollow spheres, a domain of size 2R X 2R X 2R centered the radius of the sphere. It was partitioned in N X N X N cells, where N was

Do not use 'X' but $\times$ (`\times`).

H.0.1 Computational Geodynamics Report

All the comments above apply, with additional instructions:

- report should be in LaTeX

- The document should contain your full name and student number on the first page.

- The report file should be a pdf which name contains your family name

- not more than 25 pages. If more, use appendices wisely.

- document should be structured in two main parts: FDM and FEM.

- no equations unless necessary to the discussion (still mention the equation that you are solving but refer to an external document/article/book for example).

- use lstlisting package to include code

- use `\usepackage[cm]{fullpage}` to format your document

- all codes either in appendix or in zip file (bearing your name).

- a decent introduction (half page to one page) which links the topic of this course to geosciences.

- discussion of results (stability, convergence, influence of resolution, remarks of all kinds).

- if you did not succeed in doing a particular exercise, please explain what you think the problem is, how you know it is not working, etc ...

- think about colormaps, image compression

- DEADLINE: July 16th, 2023, 23:59

I will use this table to grade your reports:

title, names, student nb																
LaTeX?																
document layout																
equations look																
equations numbered																
use of equations																
figs: caption																
figs: pixels?																
figs: correct?																
English grammar																
Typos Introduction																
methods/results																
Discussion/Conclusion																
Extra work?																
Bibliography																
code layout																
code style																
code accuracy																

Appendix I

Analytical expressions for $\mathbb{G}_{el}$

The elemental matrix $\mathbb{G}_{el}$ is given by (see Section 10.8):

$$
\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e}
\begin{pmatrix}
\partial_x \mathcal{N}_1^v & 0 & \partial_y \mathcal{N}_1^v \\
0 & \partial_y \mathcal{N}_1^v & \partial_x \mathcal{N}_1^v \\
\partial_x \mathcal{N}_2^v & 0 & \partial_y \mathcal{N}_2^v \\
0 & \partial_y \mathcal{N}_2^v & \partial_x \mathcal{N}_2^v \\
\ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots \\
\partial_x \mathcal{N}_{m_v}^v & 0 & \partial_y \mathcal{N}_{m_v}^v \\
0 & \partial_y \mathcal{N}_{m_v}^v & \partial_x \mathcal{N}_{m_v}^v
\end{pmatrix}
\cdot
\begin{pmatrix}
\mathcal{N}_1^p & \mathcal{N}_2^p & \ldots & \mathcal{N}_{m_p}^p \\
\mathcal{N}_1^p & \mathcal{N}_2^p & \ldots & \mathcal{N}_{m_p}^p \\
0 & 0 & \ldots & 0
\end{pmatrix}
d\Omega
$$

In what follows I set out to compute this elemental matrix for the reference element $[-1:1] \times [-1:1]$. All of the integrals were computed with WolframAlpha[1] since it allowed me to copy-paste the LaTeXcode directly into the website prompt area and obtain the value of these integrals.

I.0.1 $Q_1 \times P_0$ element (2d case)

For this element, $m_v = 4$ and $m_p = 1$ so $\mathbb{G}_{el}$ is a 8×1 matrix:

$$
\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e}
\begin{pmatrix}
\partial_r \mathcal{N}_1^v & 0 & \partial_s \mathcal{N}_1^v \\
0 & \partial_s \mathcal{N}_1^v & \partial_r \mathcal{N}_1^v \\
\partial_r \mathcal{N}_2^v & 0 & \partial_s \mathcal{N}_2^v \\
0 & \partial_s \mathcal{N}_2^v & \partial_r \mathcal{N}_2^v \\
\partial_r \mathcal{N}_3^v & 0 & \partial_s \mathcal{N}_3^v \\
0 & \partial_s \mathcal{N}_3^v & \partial_r \mathcal{N}_3^v \\
\partial_r \mathcal{N}_4^v & 0 & \partial_s \mathcal{N}_4^v \\
0 & \partial_s \mathcal{N}_4^v & \partial_r \mathcal{N}_4^v
\end{pmatrix}
\cdot
\begin{pmatrix}
\mathcal{N}_1^p \\
\mathcal{N}_1^p \\
0
\end{pmatrix}
d\Omega
$$

also, since $\mathcal{N}_1^p(r,s) = 1$ then

$$
\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e}
\begin{pmatrix}
\partial_r \mathcal{N}_1^v \\
\partial_s \mathcal{N}_1^v \\
\partial_r \mathcal{N}_2^v \\
\partial_s \mathcal{N}_2^v \\
\partial_r \mathcal{N}_3^v \\
\partial_s \mathcal{N}_3^v \\
\partial_r \mathcal{N}_4^v \\
\partial_s \mathcal{N}_4^v
\end{pmatrix}
d\Omega =
\begin{pmatrix}
1 \\
1 \\
-1 \\
1 \\
-1 \\
-1 \\
1 \\
-1
\end{pmatrix}
$$

---

[1] https://www.wolframalpha.com/

A macro element made of a single element makes no sense since if velocity is prescribed on all sides there is not a single velocity dof left.

We then consider the following macro-element:

```
velocity        pressure
7====8====9    .====.====.
|    |    |    |  3 |  4 |
4====5====6    .====.====.    NV=9
|    |    |    |  1 |  2 |
1====2====3    .====.====.    NP=4
```

The assembled $\mathbb{G}$ matrix is then 18×4:

$$\mathbb{G} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

After applying boundary conditions on nodes 1,2,3,4,6,7,8,9:

$$\mathbb{G} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

1313

or, defining $\tilde{\mathbb{G}}$ as the remaining matrix when the zero lines have been removed:

$$\tilde{\mathbb{G}} = \begin{pmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix}$$

The null space is of size two, spawn by the two vectors:

```
[[-0.1  0.7]
 [ 0.7  0.1]
 [ 0.7  0.1]
 [-0.1  0.7]]
```

The means that we have two pressure solutions $\vec{\mathcal{P}}_{1,2}$ (and any linear combination of these 2) which are such that $\mathbb{G} \cdot \vec{\mathcal{P}} = \vec{0}$ while we would ideally want a single one, the constant vector.

In the book (which?) the authors proceed to show that any such macro-element made of rectangles has a spurious mode.

See code `python_codes/Gel/macro_element_q1p0.py`

I.0.2 $Q_1 \times P_0$ element (3d case)

For this element, $m_v = 8$ and $m_p = 1$ so $\mathbb{G}_{el}$ is a $3 * 8 \times 1$ matrix:

$$\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e} \begin{pmatrix} \partial_r \mathcal{N}_1^\gamma & 0 & 0 & \partial_s \mathcal{N}_1^\gamma & \partial_t \mathcal{N}_1^\gamma & 0 \\ 0 & \partial_s \mathcal{N}_1^\gamma & 0 & \partial_r \mathcal{N}_1^\gamma & 0 & \partial_t \mathcal{N}_1^\gamma \\ 0 & 0 & \partial_t \mathcal{N}_1^\gamma & 0 & \partial_r \mathcal{N}_1^\gamma & \partial_s \mathcal{N}_1^\gamma \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \partial_r \mathcal{N}_8^\gamma & 0 & 0 & \partial_s \mathcal{N}_8^\gamma & \partial_t \mathcal{N}_8^\gamma & 0 \\ 0 & \partial_s \mathcal{N}_8^\gamma & 0 & \partial_r \mathcal{N}_8^\gamma & 0 & \partial_t \mathcal{N}_8^\gamma \\ 0 & 0 & \partial_t \mathcal{N}_8^\gamma & 0 & \partial_r \mathcal{N}_8^\gamma & \partial_s \mathcal{N}_8^\gamma \end{pmatrix} \cdot \begin{pmatrix} \mathcal{N}_1^p \\ \mathcal{N}_1^p \\ \mathcal{N}_1^p \\ 0 \\ 0 \\ 0 \end{pmatrix} d\Omega$$

also, since $\mathcal{N}_1^p(r,s) = 1$ then

$$\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e} \begin{pmatrix} \partial_r \mathcal{N}_1^\gamma \\ \partial_s \mathcal{N}_1^\gamma \\ \partial_t \mathcal{N}_1^\gamma \\ \partial_r \mathcal{N}_2^\gamma \\ \partial_s \mathcal{N}_2^\gamma \\ \partial_t \mathcal{N}_2^\gamma \\ \partial_r \mathcal{N}_3^\gamma \\ \partial_s \mathcal{N}_3^\gamma \\ \partial_t \mathcal{N}_3^\gamma \\ \partial_r \mathcal{N}_4^\gamma \\ \partial_s \mathcal{N}_4^\gamma \\ \partial_t \mathcal{N}_4^\gamma \\ \partial_r \mathcal{N}_5^\gamma \\ \partial_s \mathcal{N}_5^\gamma \\ \partial_t \mathcal{N}_5^\gamma \\ \partial_r \mathcal{N}_6^\gamma \\ \partial_s \mathcal{N}_6^\gamma \\ \partial_t \mathcal{N}_6^\gamma \\ \partial_r \mathcal{N}_7^\gamma \\ \partial_s \mathcal{N}_7^\gamma \\ \partial_t \mathcal{N}_7^\gamma \\ \partial_r \mathcal{N}_8^\gamma \\ \partial_s \mathcal{N}_8^\gamma \\ \partial_t \mathcal{N}_8^\gamma \end{pmatrix} d\Omega = -\begin{pmatrix} \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(-1)(1-s)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(-1)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(1-s)(-1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(+1)(1-s)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(-1)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(1-s)(-1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(+1)(1+s)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(+1)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(1+s)(-1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(-1)(1+s)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(+1)(1-t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(1+s)(-1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(-1)(1-s)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(-1)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(1-s)(+1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(+1)(1-s)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(-1)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(1-s)(+1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(+1)(1+s)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(+1)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1+r)(1+s)(+1)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(-1)(1+s)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(+1)(1+t)drdsdt \\ \int_{-1}^{+1}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{8}(1-r)(1+s)(+1)drdsdt \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

If we consider a macro-element $2 \times 2 \times 2$ of size $L_x = L_y = L_z = 4$, apply velocity b.c on the all sides we are left with

$$\tilde{G} = \begin{pmatrix} -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix}$$

Null space has dimension 5:

```
[[-0.35355339  0.35355339  0.35355339  0.35355339  0.35355339]
 [ 0.35355339 -0.35355339 -0.35355339  0.35355339  0.35355339]
 [ 0.35355339 -0.35355339  0.35355339 -0.35355339  0.35355339]
 [ 0.41990569  0.58009431  0.19336477  0.19336477 -0.19336477]
 [ 0.58009431  0.41990569 -0.19336477 -0.19336477  0.19336477]
 [ 0.19336477 -0.19336477  0.74028293  0.03317615 -0.03317615]
 [ 0.19336477 -0.19336477  0.03317615  0.74028293 -0.03317615]
 [-0.19336477  0.19336477 -0.03317615 -0.03317615  0.74028293]]
```

I.0.3 $Q_1 \times Q_1$ element (2d case)

For this element, $m_v = 4$ and $m_p = 4$ so $\mathbb{G}_{el}$ is a 8×4 matrix:

$$\mathbb{G}_{el} \;=\; -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N}\, d\Omega \tag{I.1}$$

$$= -\int_{\Omega_e} \begin{pmatrix} \partial_r\mathcal{N}_1^\gamma & 0 & \partial_s\mathcal{N}_1^\gamma \\ 0 & \partial_s\mathcal{N}_1^\gamma & \partial_r\mathcal{N}_1^\gamma \\ \partial_r\mathcal{N}_2^\gamma & 0 & \partial_s\mathcal{N}_2^\gamma \\ 0 & \partial_s\mathcal{N}_2^\gamma & \partial_r\mathcal{N}_2^\gamma \\ \partial_r\mathcal{N}_3^\gamma & 0 & \partial_s\mathcal{N}_3^\gamma \\ 0 & \partial_s\mathcal{N}_3^\gamma & \partial_r\mathcal{N}_3^\gamma \\ \partial_r\mathcal{N}_4^\gamma & 0 & \partial_s\mathcal{N}_4^\gamma \\ 0 & \partial_s\mathcal{N}_4^\gamma & \partial_r\mathcal{N}_4^\gamma \end{pmatrix} \cdot \begin{pmatrix} \mathcal{N}_1^p & \mathcal{N}_2^p & \mathcal{N}_3^p & \mathcal{N}_4^p \\ \mathcal{N}_1^p & \mathcal{N}_2^p & \mathcal{N}_3^p & \mathcal{N}_4^p \\ 0 & 0 & \dots & 0 \end{pmatrix} d\Omega \tag{I.2}$$

$$= -\int_{\Omega_e} \begin{pmatrix} \mathcal{N}_1^p\partial_r\mathcal{N}_1^\gamma & \mathcal{N}_2^p\partial_r\mathcal{N}_1^\gamma & \mathcal{N}_3^p\partial_r\mathcal{N}_1^\gamma & \mathcal{N}_4^p\partial_r\mathcal{N}_1^\gamma \\ \mathcal{N}_1^p\partial_s\mathcal{N}_1^\gamma & \mathcal{N}_2^p\partial_s\mathcal{N}_1^\gamma & \mathcal{N}_3^p\partial_s\mathcal{N}_1^\gamma & \mathcal{N}_4^p\partial_s\mathcal{N}_1^\gamma \\ \mathcal{N}_1^p\partial_r\mathcal{N}_2^\gamma & \mathcal{N}_2^p\partial_r\mathcal{N}_2^\gamma & \mathcal{N}_3^p\partial_r\mathcal{N}_2^\gamma & \mathcal{N}_4^p\partial_r\mathcal{N}_2^\gamma \\ \mathcal{N}_1^p\partial_s\mathcal{N}_2^\gamma & \mathcal{N}_2^p\partial_s\mathcal{N}_2^\gamma & \mathcal{N}_3^p\partial_s\mathcal{N}_2^\gamma & \mathcal{N}_4^p\partial_s\mathcal{N}_2^\gamma \\ \mathcal{N}_1^p\partial_r\mathcal{N}_3^\gamma & \mathcal{N}_2^p\partial_r\mathcal{N}_3^\gamma & \mathcal{N}_3^p\partial_r\mathcal{N}_3^\gamma & \mathcal{N}_4^p\partial_r\mathcal{N}_3^\gamma \\ \mathcal{N}_1^p\partial_s\mathcal{N}_3^\gamma & \mathcal{N}_2^p\partial_s\mathcal{N}_3^\gamma & \mathcal{N}_3^p\partial_s\mathcal{N}_3^\gamma & \mathcal{N}_4^p\partial_s\mathcal{N}_3^\gamma \\ \mathcal{N}_1^p\partial_r\mathcal{N}_4^\gamma & \mathcal{N}_2^p\partial_r\mathcal{N}_4^\gamma & \mathcal{N}_3^p\partial_r\mathcal{N}_4^\gamma & \mathcal{N}_4^p\partial_r\mathcal{N}_4^\gamma \\ \mathcal{N}_1^p\partial_s\mathcal{N}_4^\gamma & \mathcal{N}_2^p\partial_s\mathcal{N}_4^\gamma & \mathcal{N}_3^p\partial_s\mathcal{N}_4^\gamma & \mathcal{N}_4^p\partial_s\mathcal{N}_4^\gamma \end{pmatrix} d\Omega \tag{I.3}$$

$$\tag{I.4}$$

We have $\mathcal{N}_i^\gamma = \mathcal{N}_i^p$ with $i = 1, 2, 3, 4$, so we can drop the superscripts and we can write:

$$\mathbb{G}_{el} \;=\; -\int_{\Omega_e} \begin{pmatrix} \mathcal{N}_1\partial_r\mathcal{N}_1 & \mathcal{N}_2\partial_r\mathcal{N}_1 & \mathcal{N}_3\partial_r\mathcal{N}_1 & \mathcal{N}_4\partial_r\mathcal{N}_1 \\ \mathcal{N}_1\partial_s\mathcal{N}_1 & \mathcal{N}_2\partial_s\mathcal{N}_1 & \mathcal{N}_3\partial_s\mathcal{N}_1 & \mathcal{N}_4\partial_s\mathcal{N}_1 \\ \mathcal{N}_1\partial_r\mathcal{N}_2 & \mathcal{N}_2\partial_r\mathcal{N}_2 & \mathcal{N}_3\partial_r\mathcal{N}_2 & \mathcal{N}_4\partial_r\mathcal{N}_2 \\ \mathcal{N}_1\partial_s\mathcal{N}_2 & \mathcal{N}_2\partial_s\mathcal{N}_2 & \mathcal{N}_3\partial_s\mathcal{N}_2 & \mathcal{N}_4\partial_s\mathcal{N}_2 \\ \mathcal{N}_1\partial_r\mathcal{N}_3 & \mathcal{N}_2\partial_r\mathcal{N}_3 & \mathcal{N}_3\partial_r\mathcal{N}_3 & \mathcal{N}_4\partial_r\mathcal{N}_3 \\ \mathcal{N}_1\partial_s\mathcal{N}_3 & \mathcal{N}_2\partial_s\mathcal{N}_3 & \mathcal{N}_3\partial_s\mathcal{N}_3 & \mathcal{N}_4\partial_s\mathcal{N}_3 \\ \mathcal{N}_1\partial_r\mathcal{N}_4 & \mathcal{N}_2\partial_r\mathcal{N}_4 & \mathcal{N}_3\partial_r\mathcal{N}_4 & \mathcal{N}_4\partial_r\mathcal{N}_4 \\ \mathcal{N}_1\partial_s\mathcal{N}_4 & \mathcal{N}_2\partial_s\mathcal{N}_4 & \mathcal{N}_3\partial_s\mathcal{N}_4 & \mathcal{N}_4\partial_s\mathcal{N}_4 \end{pmatrix} d\Omega \tag{I.5}$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_r\mathcal{N}_1\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(-1)(1-s)\,dr\,ds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_s\mathcal{N}_1\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(1-r)(-1)\,dr\,ds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_r\mathcal{N}_2\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(+1)(1-s)\,dr\,ds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_s\mathcal{N}_2\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(1+r)(-1)\,dr\,ds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_r\mathcal{N}_3\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(+1)(1+s)\,dr\,ds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_s\mathcal{N}_3\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(1+r)(+1)\,dr\,ds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_r\mathcal{N}_4\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(-1)(1+s)\,dr\,ds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_1\partial_s\mathcal{N}_4\, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{4}(1-r)(+1)\,dr\,ds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_r \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(-1)(1-s)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_s \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(1-r)(-1)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_r \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(+1)(1-s)drds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_s \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(1+r)(-1)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_r \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(+1)(1+s)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_s \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(1+r)(+1)drds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_r \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(-1)(1+s)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_2 \partial_s \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{4}(1-r)(+1)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_r \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(-1)(1-s)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_s \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(1-r)(-1)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_r \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(+1)(1-s)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_s \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(1+r)(-1)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_r \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(+1)(1+s)drds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_s \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(1+r)(+1)drds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_r \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(-1)(1+s)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_3 \partial_s \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{4}(1-r)(+1)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_r \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(-1)(1-s)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_s \mathcal{N}_1 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(1-r)(-1)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_r \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(+1)(1-s)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_s \mathcal{N}_2 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(1+r)(-1)drds = -1/6$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_r \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(+1)(1+s)drds = 1/3$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_s \mathcal{N}_3 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(1+r)(+1)drds = 1/6$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_r \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(-1)(1+s)drds = -1/3$$

$$\int_{\Omega_e} \mathcal{N}_4 \partial_s \mathcal{N}_4 d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{4}(1-r)(+1)drds = 1/3$$

Putting it all together:

$$\mathbb{G}_{el} = -\begin{pmatrix} -1/3 & -1/3 & -1/6 & -1/6 \\ -1/3 & -1/6 & -1/6 & -1/3 \\ 1/3 & 1/3 & 1/6 & 1/6 \\ -1/6 & -1/3 & -1/3 & -1/6 \\ 1/6 & 1/6 & 1/3 & 1/3 \\ 1/6 & 1/3 & 1/3 & 1/6 \\ -1/6 & -1/6 & -1/3 & -1/3 \\ 1/3 & 1/6 & 1/6 & 1/3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 2 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ -2 & -2 & -1 & -1 \\ 1 & 2 & 2 & 1 \\ -1 & -1 & -2 & -2 \\ -1 & -2 & -2 & -1 \\ 1 & 1 & 2 & 2 \\ -2 & -1 & -1 & -2 \end{pmatrix} \tag{I.6}$$

I have implemented a 3×3 quadrature integration to numerically compute the matrix in the file `python_codes/Gel/programQ1Q1.py`. The code returns:

```
[[ 0.33333333   0.33333333   0.16666667   0.16666667]
 [ 0.33333333   0.16666667   0.16666667   0.33333333]
 [-0.33333333  -0.33333333  -0.16666667  -0.16666667]
 [ 0.16666667   0.33333333   0.33333333   0.16666667]
 [-0.16666667  -0.16666667  -0.33333333  -0.33333333]
 [-0.16666667  -0.33333333  -0.33333333  -0.16666667]
 [ 0.16666667   0.16666667   0.33333333   0.33333333]
 [-0.33333333  -0.16666667  -0.16666667  -0.33333333]]
```

which is indeed what we have obtained above.where???????

Similarly to the Q1P0 element, a macroelement made of a single element has zero left over vel dof after b.c. are applied on all sides, so we resort to the following macroelement:

```
velocity        pressure
7====8====9     7====8====9
|    |    |     |    |    |
4====5====6     4====5====6     NV=9
|    |    |     |    |    |
1====2====3     1====2====3     NP=9
```

After assembly we have $\mathbb{G}$ is a $ndofV * NV \times ndofP * NP = 18 * 9$ matrix:

$$
\mathbb{G} = \frac{1}{6}
\begin{pmatrix}
2 & 2 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 2 & 1 & 0 & 0 & 0 & 0 \\
-2 & 0 & 2 & -1 & 0 & 1 & 0 & 0 & 0 \\
1 & 4 & 1 & 1 & 4 & 1 & 0 & 0 & 0 \\
0 & -2 & -2 & 0 & -1 & -1 & 0 & 0 & 0 \\
0 & 1 & 2 & 0 & 1 & 2 & 0 & 0 & 0 \\
1 & 1 & 0 & 4 & 4 & 0 & 1 & 1 & 0 \\
-2 & -1 & 0 & 0 & 0 & 0 & 2 & 1 & 0 \\
-1 & 0 & 1 & -4 & 0 & 4 & -1 & 0 & 1 \\
-1 & -4 & -1 & 0 & 0 & 0 & 1 & 4 & 1 \\
0 & -1 & -1 & 0 & -4 & -4 & 0 & -1 & -1 \\
0 & -1 & -2 & 0 & 0 & 0 & 0 & 1 & 2 \\
0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 & 0 \\
0 & 0 & 0 & -2 & -1 & 0 & -2 & -1 & 0 \\
0 & 0 & 0 & -1 & 0 & 1 & -2 & 0 & 2 \\
0 & 0 & 0 & -1 & -4 & -1 & -1 & -4 & -1 \\
0 & 0 & 0 & 0 & -1 & -1 & 0 & -2 & -2 \\
0 & 0 & 0 & 0 & -1 & -2 & 0 & -1 & -2
\end{pmatrix}
$$

and after imposing boundary conditions on nodes 1,2,3,4,6,7,8,9:

$$
\mathbb{G} = \frac{1}{6} \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & -4 & 0 & 4 & -1 & 0 & 1 \\
-1 & -4 & -1 & 0 & 0 & 0 & 1 & 4 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

or,

$$
\tilde{\mathbb{G}} = \frac{1}{6} \begin{pmatrix}
-1 & 0 & 1 & -4 & 0 & 4 & -1 & 0 & 1 \\
-1 & -4 & -1 & 0 & 0 & 0 & 1 & 4 & 1
\end{pmatrix}
$$

When passed to *null_space* as argument it returns the following nullspace:

```
[[ 1.47619e-01 -6.57142e-01 -7.40148e-18  6.57142e-01 -1.47619e-01  6.66666e-02  1.80952e-
 [-1.90476e-01  9.52380e-02 -7.40148e-17 -9.52380e-02  1.90476e-01  6.66666e-01  1.42857e-
 [ 9.57142e-01  1.04761e-01 -7.40148e-18 -1.04761e-01  4.28571e-02  6.66666e-02 -9.52380e-
 [ 9.52380e-02  6.19047e-01  2.34780e-34  3.80952e-01 -9.52380e-02 -2.11471e-18  9.52380e-
 [-8.45884e-18  4.22942e-18  1.00000e+00 -4.22942e-18  8.45884e-18  2.96059e-17  6.34413e-
 [-9.52380e-02  3.80952e-01  5.35591e-34  6.19047e-01  9.52380e-02 -4.82418e-18 -9.52380e-
 [ 4.28571e-02 -1.04761e-01  7.40148e-18  1.04761e-01  9.57142e-01 -6.66666e-02  9.52380e-
 [ 7.61904e-02 -3.80952e-02  2.96059e-17  3.80952e-02 -7.61904e-02  7.33333e-01 -5.71428e-
 [-4.76190e-03  8.57142e-02  7.40148e-18 -8.57142e-02  4.76190e-03 -6.66666e-02  9.61904e-
```

which is very bad: the dimension of the nullspace is 9!

See code `python_codes/Gel/macro_element_q1q1.py`

I.0.4 $Q_2 \times Q_1$ element

$$\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega$$

$$= -\int_{\Omega_e} \begin{pmatrix} \partial_x \mathcal{N}_1^\gamma & 0 & \partial_y \mathcal{N}_1^\gamma \\ 0 & \partial_y \mathcal{N}_1^\gamma & \partial_x \mathcal{N}_1^\gamma \\ \partial_x \mathcal{N}_2^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma \\ 0 & \partial_y \mathcal{N}_2^\gamma & \partial_x \mathcal{N}_2^\gamma \\ \partial_x \mathcal{N}_3^\gamma & 0 & \partial_y \mathcal{N}_1^\gamma \\ 0 & \partial_y \mathcal{N}_3^\gamma & \partial_x \mathcal{N}_1^\gamma \\ \partial_x \mathcal{N}_4^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma \\ 0 & \partial_y \mathcal{N}_4^\gamma & \partial_x \mathcal{N}_2^\gamma \\ \partial_x \mathcal{N}_5^\gamma & 0 & \partial_y \mathcal{N}_1^\gamma \\ 0 & \partial_y \mathcal{N}_5^\gamma & \partial_x \mathcal{N}_1^\gamma \\ \partial_x \mathcal{N}_6^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma \\ 0 & \partial_y \mathcal{N}_6^\gamma & \partial_x \mathcal{N}_2^\gamma \\ \partial_x \mathcal{N}_7^\gamma & 0 & \partial_y \mathcal{N}_1^\gamma \\ 0 & \partial_y \mathcal{N}_7^\gamma & \partial_x \mathcal{N}_1^\gamma \\ \partial_x \mathcal{N}_8^\gamma & 0 & \partial_y \mathcal{N}_2^\gamma \\ 0 & \partial_y \mathcal{N}_8^\gamma & \partial_x \mathcal{N}_2^\gamma \\ \partial_x \mathcal{N}_9^\gamma & 0 & \partial_y \mathcal{N}_1^\gamma \\ 0 & \partial_y \mathcal{N}_9^\gamma & \partial_x \mathcal{N}_1^\gamma \end{pmatrix} \cdot \begin{pmatrix} \mathcal{N}_1^p & \mathcal{N}_2^p & \mathcal{N}_3^p & \mathcal{N}_4^p \\ \mathcal{N}_1^p & \mathcal{N}_2^p & \mathcal{N}_3^p & \mathcal{N}_4^p \\ 0 & 0 & \dots & 0 \end{pmatrix} d\Omega$$

$$= -\int_{\Omega_e} \begin{pmatrix} \mathcal{N}_1^p \partial_r \mathcal{N}_1^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_1^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_1^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_1^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_1^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_1^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_1^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_1^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_2^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_2^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_2^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_2^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_2^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_2^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_2^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_2^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_3^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_3^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_3^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_3^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_3^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_3^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_3^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_3^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_4^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_4^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_4^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_4^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_4^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_4^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_4^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_4^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_5^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_5^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_5^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_5^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_5^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_5^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_5^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_5^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_6^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_6^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_6^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_6^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_6^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_6^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_6^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_6^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_7^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_7^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_7^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_7^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_7^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_7^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_7^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_7^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_8^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_8^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_8^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_8^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_8^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_8^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_8^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_8^\gamma \\ \mathcal{N}_1^p \partial_r \mathcal{N}_9^\gamma & \mathcal{N}_2^p \partial_r \mathcal{N}_9^\gamma & \mathcal{N}_3^p \partial_r \mathcal{N}_9^\gamma & \mathcal{N}_4^p \partial_r \mathcal{N}_9^\gamma \\ \mathcal{N}_1^p \partial_s \mathcal{N}_9^\gamma & \mathcal{N}_2^p \partial_s \mathcal{N}_9^\gamma & \mathcal{N}_3^p \partial_s \mathcal{N}_9^\gamma & \mathcal{N}_4^p \partial_s \mathcal{N}_9^\gamma \end{pmatrix} d\Omega \qquad (\text{I.7})$$

$$\int_{\Omega_e} \mathcal{N}_1^p \partial_r \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{2}(2r-1)\frac{1}{2}s(s-1)drds = -5/18$$

$$\int_{\Omega_e} \mathcal{N}_2^p \partial_s \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)\frac{1}{2}r(r-1)\frac{1}{2}(2s-1)drds = -5/18$$

$$\int_{\Omega_e} \mathcal{N}_3^p \partial_r \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{2}(2r-1)\frac{1}{2}s(s-1)drds = -1/18$$

$$\int_{\Omega_e} \mathcal{N}_4^p \partial_s \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)\frac{1}{2}r(r-1)\frac{1}{2}(2s-1)drds = 0$$

$$\int_{\Omega_e} \mathcal{N}_1^p \partial_r \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{2}(2r-1)\frac{1}{2}s(s-1)drds = 0$$

$$\int_{\Omega_e} \mathcal{N}_2^p \partial_s \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)\frac{1}{2}r(r-1)\frac{1}{2}(2s-1)drds = 0$$

$$\int_{\Omega_e} \mathcal{N}_3^p \partial_r \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{2}(2r-1)\frac{1}{2}s(s-1)drds = 0$$

$$\int_{\Omega_e} \mathcal{N}_4^p \partial_s \mathcal{N}_1^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)\frac{1}{2}r(r-1)\frac{1}{2}(2s-1)drds = -1/18$$

... same procedure for 1,2,3,4,5,6,7...

$$\int_{\Omega_e} \mathcal{N}_1^p \partial_r \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)(-2r)(1-s^2)drds = 4/9$$

$$\int_{\Omega_e} \mathcal{N}_2^p \partial_s \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)(1-r^2)(-2s)drds = 4/9$$

$$\int_{\Omega_e} \mathcal{N}_3^p \partial_r \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)(-2r)(1-s^2)drds = -4/9$$

$$\int_{\Omega_e} \mathcal{N}_4^p \partial_s \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)(1-r^2)(-2s)drds = 4/9$$

$$\int_{\Omega_e} \mathcal{N}_1^p \partial_r \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)(-2r)(1-s^2)drds = -4/9$$

$$\int_{\Omega_e} \mathcal{N}_2^p \partial_s \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)(1-r^2)(-2s)drds = -4/9$$

$$\int_{\Omega_e} \mathcal{N}_3^p \partial_r \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)(-2r)(1-s^2)drds = 4/9$$

$$\int_{\Omega_e} \mathcal{N}_4^p \partial_s \mathcal{N}_9^\gamma \, d\Omega = \int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)(1-r^2)(-2s)drds = -4/9$$

We obtain

$$\tilde{\mathbb{G}}_{el} = -\int_{\Omega_e} \begin{pmatrix} N_1^p \partial_r N_9^\gamma & N_2^p \partial_r N_9^\gamma & N_3^p \partial_r N_9^\gamma & N_4^p \partial_r N_9^\gamma \\ N_1^p \partial_s N_9^\gamma & N_2^p \partial_s N_9^\gamma & N_3^p \partial_s N_9^\gamma & N_4^p \partial_s N_9^\gamma \end{pmatrix} d\Omega = -\frac{4}{9}\begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}$$

which is identical to Eq.(3.53) of Elman, Silvester, and Wathen [394]. I have implemented a 3x3 quadrature integration to numerically compute the matrix in the file `python_codes/Gel/programQ2Q1.py`:

```
[[ 2.77777778e-01  5.55555556e-02  1.73472348e-18  3.68628739e-18]
 [ 2.77777778e-01 -1.29020059e-17 -3.46944695e-18  5.55555556e-02]
 [-5.55555556e-02 -2.77777778e-01 -2.77555756e-17 -3.03576608e-18]
 [-2.19008839e-17  2.77777778e-01  5.55555556e-02 -4.33680869e-18]
 [-3.46944695e-18 -1.38777878e-17 -2.77777778e-01 -5.55555556e-02]
 [ 2.60208521e-18 -5.55555556e-02 -2.77777778e-01  6.93889390e-18]
 [ 4.01154804e-18  4.33680869e-18  5.55555556e-02  2.77777778e-01]
 [-5.55555556e-02  6.07153217e-18  2.08166817e-17 -2.77777778e-01]
 [-2.22222222e-01  2.22222222e-01  0.00000000e+00 -8.67361738e-19]
 [ 5.55555556e-01  5.55555556e-01  1.11111111e-01  1.11111111e-01]
 [-1.11111111e-01 -5.55555556e-01 -5.55555556e-01 -1.11111111e-01]
 [-8.67361738e-19 -2.22222222e-01  2.22222222e-01  6.93889390e-18]
 [-8.67361738e-18 -6.93889390e-18  2.22222222e-01 -2.22222222e-01]
 [-1.11111111e-01 -1.11111111e-01 -5.55555556e-01 -5.55555556e-01]
 [ 5.55555556e-01  1.11111111e-01  1.11111111e-01  5.55555556e-01]
 [-2.22222222e-01 -5.20417043e-18  6.93889390e-18  2.22222222e-01]
 [-4.44444444e-01  4.44444444e-01  4.44444444e-01 -4.44444444e-01]
 [-4.44444444e-01 -4.44444444e-01  4.44444444e-01  4.44444444e-01]]
```

or,

$$
\mathbb{G}_{el} = \frac{1}{18}
\begin{pmatrix}
5 & 1 & 0 & 0 \\
5 & 0 & 0 & 1 \\
-1 & -5 & 0 & 0 \\
0 & 5 & 1 & 0 \\
0 & 0 & -5 & -1 \\
0 & -1 & -5 & 0 \\
0 & 0 & 1 & 5 \\
-1 & 0 & 0 & -5 \\
-4 & 4 & 0 & 0 \\
10 & 10 & 2 & 2 \\
-2 & -10 & -10 & -2 \\
0 & -4 & 4 & 0 \\
0 & 0 & 4 & -4 \\
-2 & -2 & -10 & -10 \\
10 & 2 & 2 & 10 \\
-4 & 0 & 0 & 4 \\
-8 & 8 & 8 & -8 \\
-8 & -8 & 8 & 8
\end{pmatrix}
$$

Reading Elman, Silvester, and Wathen [394] we see that this element is stable but patches of even and odd elements actually are needed to establish the stability of the element.

I.0.5 $Q_1^+ \times Q_1$ element

I need to revisit this section, add bN, etc ....

For the quadrilateral MINI element, $m_v = 5$ and $m_p = 4$ so $\mathbb{G}_{el}$ is a 10×4 matrix:

$$
\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega \tag{I.8}
$$

$$
= -\int_{\Omega_e}
\begin{pmatrix}
\partial_r N_1^v & 0 & \partial_s N_1^v \\
0 & \partial_s N_1^v & \partial_r N_1^v \\
\partial_r N_2^v & 0 & \partial_s N_2^v \\
0 & \partial_s N_2^v & \partial_r N_2^v \\
\partial_r N_3^v & 0 & \partial_s N_3^v \\
0 & \partial_s N_3^v & \partial_r N_3^v \\
\partial_r N_4^v & 0 & \partial_s N_4^v \\
0 & \partial_s N_4^v & \partial_r N_4^v \\
\partial_r N_5^v & 0 & \partial_s N_5^v \\
0 & \partial_s N_5^v & \partial_r N_5^v
\end{pmatrix}
\cdot
\begin{pmatrix}
N_1^p & N_2^p & N_3^p & N_4^p \\
N_1^p & N_2^p & N_3^p & N_4^p \\
0 & 0 & \ldots & 0
\end{pmatrix}
d\Omega \tag{I.9}
$$

$$
= -\int_{\Omega_e}
\begin{pmatrix}
N_1^p \partial_r N_1^v & N_2^p \partial_r N_1^v & N_3^p \partial_r N_1^v & N_4^p \partial_r N_1^v \\
N_1^p \partial_s N_1^v & N_2^p \partial_s N_1^v & N_3^p \partial_s N_1^v & N_4^p \partial_s N_1^v \\
N_1^p \partial_r N_2^v & N_2^p \partial_r N_2^v & N_3^p \partial_r N_2^v & N_4^p \partial_r N_2^v \\
N_1^p \partial_s N_2^v & N_2^p \partial_s N_2^v & N_3^p \partial_s N_2^v & N_4^p \partial_s N_2^v \\
N_1^p \partial_r N_3^v & N_2^p \partial_r N_3^v & N_3^p \partial_r N_3^v & N_4^p \partial_r N_3^v \\
N_1^p \partial_s N_3^v & N_2^p \partial_s N_3^v & N_3^p \partial_s N_3^v & N_4^p \partial_s N_3^v \\
N_1^p \partial_r N_4^v & N_2^p \partial_r N_4^v & N_3^p \partial_r N_4^v & N_4^p \partial_r N_4^v \\
N_1^p \partial_s N_4^v & N_2^p \partial_s N_4^v & N_3^p \partial_s N_4^v & N_4^p \partial_s N_4^v \\
N_1^p \partial_r N_5^v & N_2^p \partial_r N_5^v & N_3^p \partial_r N_5^v & N_4^p \partial_r N_5^v \\
N_1^p \partial_s N_5^v & N_2^p \partial_s N_5^v & N_3^p \partial_s N_5^v & N_4^p \partial_s N_5^v
\end{pmatrix}
d\Omega \tag{I.10}
$$

We have :

$$N_1^v = N_1^p - \frac{1}{4}b(r,s) \tag{I.11}$$

$$N_2^v = N_2^p - \frac{1}{4}b(r,s) \tag{I.12}$$

$$N_3^v = N_3^p - \frac{1}{4}b(r,s) \tag{I.13}$$

$$N_4^v = N_4^p - \frac{1}{4}b(r,s) \tag{I.14}$$

$$N_5^v = b(r,s) \tag{I.15}$$

so that (once again I drop the superscripts)

$$\mathbb{G}_{el} = -\int_{\Omega_e}
\begin{pmatrix}
N_1\partial_r N_1 & N_2\partial_r N_1 & N_3\partial_r N_1 & N_4\partial_r N_1 \\
N_1\partial_s N_1 & N_2\partial_s N_1 & N_3\partial_s N_1 & N_4\partial_s N_1 \\
N_1\partial_r N_2 & N_2\partial_r N_2 & N_3\partial_r N_2 & N_4\partial_r N_2 \\
N_1\partial_s N_2 & N_2\partial_s N_2 & N_3\partial_s N_2 & N_4\partial_s N_2 \\
N_1\partial_r N_3 & N_2\partial_r N_3 & N_3\partial_r N_3 & N_4\partial_r N_3 \\
N_1\partial_s N_3 & N_2\partial_s N_3 & N_3\partial_s N_3 & N_4\partial_s N_3 \\
N_1\partial_r N_4 & N_2\partial_r N_4 & N_3\partial_r N_4 & N_4\partial_r N_4 \\
N_1\partial_s N_4 & N_2\partial_s N_4 & N_3\partial_s N_4 & N_4\partial_s N_4 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix} d\Omega + \frac{1}{4}\int_{\Omega_e}
\begin{pmatrix}
N_1\partial_r b & N_2\partial_r b & N_3\partial_r b & N_4\partial_r b \\
N_1\partial_s b & N_2\partial_s b & N_3\partial_s b & N_4\partial_s b \\
N_1\partial_r b & N_2\partial_r b & N_3\partial_r b & N_4\partial_r b \\
N_1\partial_s b & N_2\partial_s b & N_3\partial_s b & N_4\partial_s b \\
N_1\partial_r b & N_2\partial_r b & N_3\partial_r b & N_4\partial_r b \\
N_1\partial_s b & N_2\partial_s b & N_3\partial_s b & N_4\partial_s b \\
N_1\partial_r b & N_2\partial_r b & N_3\partial_r b & N_4\partial_r b \\
N_1\partial_s b & N_2\partial_s b & N_3\partial_s b & N_4\partial_s b \\
-4N_1\partial_r b & -4N_2\partial_r b & -4N_3\partial_r b & -4N_4\partial_r b \\
-4N_1\partial_s b & -4N_2\partial_s b & -4N_3\partial_s b & -4N_4\partial_s b
\end{pmatrix}$$

The matrix which only contains N_i functions is in fact the $\mathbb{G}_{el}$ matrix for standard $Q_1 \times Q_1$ elements as we have seen in the previous section so we need not recompute it.

Bubble function 1

Let us now assume the bubble is bubble 1:

$$b_1(r,s) = (1-r)(1-s)(1-r^2)(1-s^2)$$
$$\partial_r b_1(r,s) = (3r^2 - 2r - 1)(1-s)(1-s^2)$$
$$\partial_s b_1(r,s) = (3s^2 - 2s - 1)(1-r)(1-r^2)$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_1\partial_r b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)(3r^2-2r-1)(1-s)(1-s^2)drds = 2/15$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_1\partial_s b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1-s)(3s^2-2s-1)(1-r)(1-r^2)drds = 2/15$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_2\partial_r b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)(3r^2-2r-1)(1-s)(1-s^2)drds = -2/15$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_2\partial_s b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1-s)(3s^2-2s-1)(1-r)(1-r^2)drds = 4/45$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_3\partial_r b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)(3r^2-2r-1)(1-s)(1-s^2)drds = -4/45$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_3\partial_s b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1+r)(1+s)(3s^2-2s-1)(1-r)(1-r^2)drds = -4/45$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_4\partial_r b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)(3r^2-2r-1)(1-s)(1-s^2)drds = 4/45$$

$$\frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} N_4\partial_s b_1\, drds = \frac{1}{4}\int_{-1}^{+1}\int_{-1}^{+1} \frac{1}{4}(1-r)(1+s)(3s^2-2s-1)(1-r)(1-r^2)drds = -2/15$$

$$\mathbb{G}_{el} \;=\; - \begin{pmatrix} -1/3 & -1/3 & -1/6 & -1/6 \\ -1/3 & -1/6 & -1/6 & -1/3 \\ 1/3 & 1/3 & 1/6 & 1/6 \\ -1/6 & -1/3 & -1/3 & -1/6 \\ 1/6 & 1/6 & 1/3 & 1/3 \\ 1/6 & 1/3 & 1/3 & 1/6 \\ -1/6 & -1/6 & -1/3 & -1/3 \\ 1/3 & 1/6 & 1/6 & 1/3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 2/15 & -2/15 & -4/45 & 4/45 \\ 2/15 & 4/45 & -4/45 & -2/15 \\ 2/15 & -2/15 & -4/45 & 4/45 \\ 2/15 & 4/45 & -4/45 & -2/15 \\ 2/15 & -2/15 & -4/45 & 4/45 \\ 2/15 & 4/45 & -4/45 & -2/15 \\ 2/15 & -2/15 & -4/45 & 4/45 \\ 2/15 & 4/45 & -4/45 & -2/15 \\ -8/15 & 8/15 & 16/45 & -16/45 \\ -8/15 & -16/45 & 16/45 & 8/15 \end{pmatrix} \quad \text{(I.16)}$$

I have implemented a 3x3 quadrature integration to numerically compute the matrix in the file `python_codes/Gel/programQ1pQ1.py`. The code returns:

```
[[ 0.46666667  0.2          0.07777778  0.25555556]
 [ 0.46666667  0.25555556   0.07777778  0.2         ]
 [-0.2         -0.46666667  -0.25555556 -0.07777778]
 [ 0.3          0.42222222   0.24444444  0.03333333]
 [-0.03333333  -0.3         -0.42222222 -0.24444444]
 [-0.03333333  -0.24444444  -0.42222222 -0.3         ]
 [ 0.3          0.03333333   0.24444444  0.42222222]
 [-0.2         -0.07777778  -0.25555556 -0.46666667]
 [-0.53333333   0.53333333   0.35555556 -0.35555556]
 [-0.53333333  -0.35555556   0.35555556  0.53333333]]
```

which is indeed what we have obtained above. It can be rewritten

$$\mathbb{G}_{el} = \frac{1}{90} \begin{pmatrix} 42 & 18 & 7 & 23 \\ 42 & 23 & 7 & 18 \\ -18 & -42 & -23 & -7 \\ 27 & 38 & 22 & 3 \\ -3 & -27 & -38 & -22 \\ -3 & -22 & -38 & -27 \\ 27 & 3 & 22 & 38 \\ -18 & -7 & -23 & -42 \\ -48 & 48 & 32 & -32 \\ -48 & -32 & 32 & 48 \end{pmatrix}$$

Let us now build a macroelement of size LxxLy=4x4 made of 2x2 elements. Each element has a Gel like the one above since they are of size 2x2:

```
velocity        pressure
7====8====9     7====8====9
| 12 | 13 |     |    |    |
4====5====6     4====5====6    NV=13
| 10 | 11 |     |    |    |
1====2====3     1====2====3    NP=9
```

I am here following the approach by Lamichhane [784] but I am not sure why he did not use a single element macro-element? Probably because when applying bc on all four nodes of a single element, the left over matrix $\tilde{G}_{el}$ is composed of the last two rows of $\mathbb{G}_{el}$ and this has a nullspace of dimension 2.

After assembly we have $\mathbb{G}$ is a $ndofV * NV \times ndofP * NP = 26 * 9$ matrix:

$$\mathbb{G} = \frac{1}{90} \begin{pmatrix}
42 & 18 & 0 & 23 & 7 & 0 & 0 & 0 & 0 \\
42 & 23 & 0 & 18 & 7 & 0 & 0 & 0 & 0 \\
-18 & 0 & 18 & -7 & 0 & 7 & 0 & 0 & 0 \\
27 & 80 & 23 & 3 & 40 & 7 & 0 & 0 & 0 \\
0 & -18 & -42 & 0 & -7 & -23 & 0 & 0 & 0 \\
0 & 27 & 38 & 0 & 3 & 22 & 0 & 0 & 0 \\
27 & 3 & 0 & 80 & 40 & 0 & 23 & 7 & 0 \\
-18 & -7 & 0 & 0 & 0 & 0 & 18 & 7 & 0 \\
-3 & 0 & 3 & -40 & 0 & 40 & -7 & 0 & 7 \\
-3 & -40 & -7 & 0 & 0 & 0 & 3 & 40 & 7 \\
0 & -3 & -27 & 0 & -40 & -80 & 0 & -7 & -23 \\
0 & -3 & -22 & 0 & 0 & 0 & 0 & 3 & 22 \\
0 & 0 & 0 & 27 & 3 & 0 & 38 & 22 & 0 \\
0 & 0 & 0 & -18 & -7 & 0 & -42 & -23 & 0 \\
0 & 0 & 0 & -3 & 0 & 3 & -22 & 0 & 22 \\
0 & 0 & 0 & -3 & -40 & -7 & -27 & -80 & -23 \\
0 & 0 & 0 & 0 & -3 & -27 & 0 & -22 & -38 \\
0 & 0 & 0 & 0 & -3 & -22 & 0 & -27 & -38 \\
-48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 & 0 \\
-48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 & 0 \\
0 & -48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 \\
0 & -48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 \\
0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 & 0 \\
0 & 0 & 0 & -48 & -32 & 0 & 48 & 32 & 0 \\
0 & 0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 \\
0 & 0 & 0 & 0 & -48 & -32 & 0 & 48 & 32
\end{pmatrix}$$

After boundary conditions on Vnodes 1,2,3,4,6,7,8,9, the matrix G looks like:

$$\mathbb{G} = \frac{1}{90}
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-3 & 0 & 3 & -40 & 0 & 40 & -7 & 0 & 7 \\
-3 & -40 & -7 & 0 & 0 & 0 & 3 & 40 & 7 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 & 0 \\
-48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 & 0 \\
0 & -48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 \\
0 & -48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 \\
0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 & 0 \\
0 & 0 & 0 & -48 & -32 & 0 & 48 & 32 & 0 \\
0 & 0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 \\
0 & 0 & 0 & 0 & -48 & -32 & 0 & 48 & 32
\end{pmatrix}$$

or, removing the lines with only zeros, we arrive at a 10×9 matrix as in Lamichhane [784]:

$$\tilde{\mathbb{G}} = \frac{1}{90}
\begin{pmatrix}
-3 & 0 & 3 & -40 & 0 & 40 & -7 & 0 & 7 \\
-3 & -40 & -7 & 0 & 0 & 0 & 3 & 40 & 7 \\
-48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 & 0 \\
-48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 & 0 \\
0 & -48 & 48 & 0 & -32 & 32 & 0 & 0 & 0 \\
0 & -48 & -32 & 0 & 48 & 32 & 0 & 0 & 0 \\
0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 & 0 \\
0 & 0 & 0 & -48 & -32 & 0 & 48 & 32 & 0 \\
0 & 0 & 0 & 0 & -48 & 48 & 0 & -32 & 32 \\
0 & 0 & 0 & 0 & -48 & -32 & 0 & 48 & 32
\end{pmatrix}$$

This matrix is then passed as argument to the *null_space* function which returns a single vector such that $ker(\tilde{\mathbb{G}}) = (1, 1, 1, 1, 1, 1, 1, 1, 1)$.

It must be said that the matrix above contains similar values as the one in [784] as well as the same number of nonzeros. Similarities: 16 times $\pm 48/90 = \pm 8/15 = 2 \cdot 4/15$ and 16 times $\pm 32/90 = \pm 16/45 = 2 \cdot 8/45$ as in the paper (aside from scaling factor 2). However the 12 remaining values differ ?

Bubble function 2

When looking at bubble 2 with $\beta = 1/4$, we get

$$\mathbb{G}_{el} = \frac{1}{180} \begin{pmatrix} 79 & 41 & 9 & 51 \\ 79 & 51 & 9 & 41 \\ -41 & -79 & -51 & -9 \\ 49 & 81 & 39 & 11 \\ -11 & -49 & -81 & -39 \\ -11 & -39 & -81 & -49 \\ 49 & 11 & 39 & 81 \\ -41 & -9 & -51 & -79 \\ -76 & 76 & 84 & -84 \\ -76 & -84 & 84 & 76 \end{pmatrix}$$

After assembly we have $\mathbb{G}$ is a $ndofV * NV \times ndofP * NP = 26 * 9$ matrix:

$$\mathbb{G} = \frac{1}{180} \begin{pmatrix}
79 & 41 & 0 & 51 & 9 & 0 & 0 & 0 & 0 \\
79 & 51 & 0 & 41 & 9 & 0 & 0 & 0 & 0 \\
-41 & 0 & 41 & -9 & 0 & 9 & 0 & 0 & 0 \\
49 & 160 & 51 & 11 & 80 & 9 & 0 & 0 & 0 \\
0 & -41 & -79 & 0 & -9 & -51 & 0 & 0 & 0 \\
0 & 49 & 81 & 0 & 11 & 39 & 0 & 0 & 0 \\
49 & 11 & 0 & 160 & 80 & 0 & 51 & 9 & 0 \\
-41 & -9 & 0 & 0 & 0 & 0 & 41 & 9 & 0 \\
-11 & 0 & 11 & -80 & 0 & 80 & -9 & 0 & 9 \\
-11 & -80 & -9 & 0 & 0 & 0 & 11 & 80 & 9 \\
0 & -11 & -49 & 0 & -80 & -160 & 0 & -9 & -51 \\
0 & -11 & -39 & 0 & 0 & 0 & 0 & 11 & 39 \\
0 & 0 & 0 & 49 & 11 & 0 & 81 & 39 & 0 \\
0 & 0 & 0 & -41 & -9 & 0 & -79 & -51 & 0 \\
0 & 0 & 0 & -11 & 0 & 11 & -39 & 0 & 39 \\
0 & 0 & 0 & -11 & -80 & -9 & -49 & -160 & -51 \\
0 & 0 & 0 & 0 & -11 & -49 & 0 & -39 & -81 \\
0 & 0 & 0 & 0 & -11 & -39 & 0 & -49 & -81 \\
-76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 & 0 \\
-76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 & 0 \\
0 & -76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 \\
0 & -76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 \\
0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 & 0 \\
0 & 0 & 0 & -76 & -84 & 0 & 76 & 84 & 0 \\
0 & 0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 \\
0 & 0 & 0 & 0 & -76 & -84 & 0 & 76 & 84
\end{pmatrix}$$

After boundary conditions are applied:

$$\mathbb{G} = \frac{1}{180}\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-11 & 0 & 11 & -80 & 0 & 80 & -9 & 0 & 9 \\
-11 & -80 & -9 & 0 & 0 & 0 & 11 & 80 & 9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 & 0 \\
-76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 & 0 \\
0 & -76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 \\
0 & -76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 \\
0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 & 0 \\
0 & 0 & 0 & -76 & -84 & 0 & 76 & 84 & 0 \\
0 & 0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 \\
0 & 0 & 0 & 0 & -76 & -84 & 0 & 76 & 84
\end{pmatrix}$$

or,

$$\tilde{\mathbb{G}} = \frac{1}{180}\begin{pmatrix}
-11 & 0 & 11 & -80 & 0 & 80 & -9 & 0 & 9 \\
-11 & -80 & -9 & 0 & 0 & 0 & 11 & 80 & 9 \\
-76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 & 0 \\
-76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 & 0 \\
0 & -76 & 76 & 0 & -84 & 84 & 0 & 0 & 0 \\
0 & -76 & -84 & 0 & 76 & 84 & 0 & 0 & 0 \\
0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 & 0 \\
0 & 0 & 0 & -76 & -84 & 0 & 76 & 84 & 0 \\
0 & 0 & 0 & 0 & -76 & 76 & 0 & -84 & 84 \\
0 & 0 & 0 & 0 & -76 & -84 & 0 & 76 & 84
\end{pmatrix}$$

We make the same observation as for bubble 1: when this matrix is passed as argument to the *null_space* function, it returns a single vector such that $ker(\tilde{\mathbb{G}}) = (1, 1, 1, 1, 1, 1, 1, 1, 1)$.

Special case: $\beta = 0$ The bubble is then

$$b(r, s) = (1 - r^2)(1 - s^2)$$

We repeat the same process and arrive at

$$\mathbb{G}_{el} = \frac{1}{18} \begin{pmatrix} 8 & 4 & 1 & 5 \\ 8 & 5 & 1 & 4 \\ -4 & -8 & -5 & -1 \\ 5 & 8 & 4 & 1 \\ -1 & -5 & -8 & -4 \\ -1 & -4 & -8 & -5 \\ 5 & 1 & 4 & 8 \\ -4 & -1 & -5 & -8 \\ -8 & 8 & 8 & -8 \\ -8 & -8 & 8 & 8 \end{pmatrix}$$

and

$$\tilde{\mathbb{G}} = \frac{1}{18} \begin{pmatrix} -1 & 0 & 1 & -8 & 0 & 8 & -1 & 0 & 1 \\ -1 & -8 & -1 & 0 & 0 & 0 & 1 & 8 & 1 \\ -8 & 8 & 0 & -8 & 8 & 0 & 0 & 0 & 0 \\ -8 & -8 & 0 & 8 & 8 & 0 & 0 & 0 & 0 \\ 0 & -8 & 8 & 0 & -8 & 8 & 0 & 0 & 0 \\ 0 & -8 & -8 & 0 & 8 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & -8 & 8 & 0 & -8 & 8 & 0 \\ 0 & 0 & 0 & -8 & -8 & 0 & 8 & 8 & 0 \\ 0 & 0 & 0 & 0 & -8 & 8 & 0 & -8 & 8 \\ 0 & 0 & 0 & 0 & -8 & -8 & 0 & 8 & 8 \end{pmatrix}$$

Finally the *null_space* function returns:

```
[[ 0.27870965 -0.34974409]
 [ 0.39102578  0.31160686]
 [ 0.27870965 -0.34974409]
 [ 0.39102578  0.31160686]
 [ 0.27870965 -0.34974409]
 [ 0.39102578  0.31160686]
 [ 0.27870965 -0.34974409]
 [ 0.39102578  0.31160686]
 [ 0.27870965 -0.34974409]]
```

i.e. the null space has dimension 2, so that the element is then not stable.

See code `python_codes/Gel/macro_element_q1pq1.py`

I.0.6 $Q_1^+ \times Q_1$ element in 3D

For the quadrilateral MINI element, $m_v = 9$ and $m_p = 8$ so $\mathbb{G}_{el}$ is a 27×8 matrix (obtained with 10^3 quadrature points, no difference with 6^3 points).

$$\mathbb{G}_{el} = -\int_{\Omega_e} \boldsymbol{B}^T \cdot \boldsymbol{N} d\Omega = -\int_{\Omega_e} \begin{pmatrix} \partial_r N_1^\gamma & 0 & 0 & \partial_s N_1^\gamma & \partial_t N_1^\gamma & 0 \\ 0 & \partial_s N_1^\gamma & 0 & \partial_r N_1^\gamma & 0 & \partial_t N_1^\gamma \\ 0 & 0 & \partial_t N_1^\gamma & 0 & \partial_r N_1^\gamma & \partial_s N_1^\gamma \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \partial_r N_{27}^\gamma & 0 & 0 & \partial_s N_{27}^\gamma & \partial_t N_{27}^\gamma & 0 \\ 0 & \partial_s N_{27}^\gamma & 0 & \partial_r N_{27}^\gamma & 0 & \partial_t N_{27}^\gamma \\ 0 & 0 & \partial_t N_{27}^\gamma & 0 & \partial_r N_{27}^\gamma & \partial_s N_{27}^\gamma \end{pmatrix} \cdot \begin{pmatrix} N_1^p & N_2^p & \dots & N_8^p \\ N_1^p & N_2^p & \dots & N_8^p \\ N_1^p & N_2^p & \dots & N_8^p \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix} d\Omega$$

$$\mathbb{G}_{el} = -\int_{\Omega_e} \begin{pmatrix} N_1^p \partial_r N_1^\gamma & N_2^p \partial_r N_1^\gamma & N_3^p \partial_r N_1^\gamma & N_4^p \partial_r N_1^\gamma & N_5^p \partial_r N_6^\gamma & N_2^p \partial_r N_1^\gamma & N_7^p \partial_r N_1^\gamma & N_8^p \partial_r N_1^\gamma \\ N_1^p \partial_s N_1^\gamma & N_2^p \partial_s N_1^\gamma & N_3^p \partial_s N_1^\gamma & N_4^p \partial_s N_1^\gamma & N_5^p \partial_s N_6^\gamma & N_2^p \partial_s N_1^\gamma & N_7^p \partial_s N_1^\gamma & N_8^p \partial_s N_1^\gamma \\ N_1^p \partial_t N_1^\gamma & N_2^p \partial_t N_1^\gamma & N_3^p \partial_t N_1^\gamma & N_4^p \partial_t N_1^\gamma & N_5^p \partial_t N_6^\gamma & N_2^p \partial_t N_1^\gamma & N_7^p \partial_t N_1^\gamma & N_8^p \partial_t N_1^\gamma \\ N_1^p \partial_r N_2^\gamma & N_2^p \partial_r N_2^\gamma & N_3^p \partial_r N_2^\gamma & N_4^p \partial_r N_2^\gamma & N_5^p \partial_r N_2^\gamma & N_6^p \partial_r N_2^\gamma & N_7^p \partial_r N_2^\gamma & N_8^p \partial_r N_2^\gamma \\ N_1^p \partial_s N_2^\gamma & N_2^p \partial_s N_2^\gamma & N_3^p \partial_s N_2^\gamma & N_4^p \partial_s N_2^\gamma & N_5^p \partial_s N_2^\gamma & N_6^p \partial_s N_2^\gamma & N_7^p \partial_s N_2^\gamma & N_8^p \partial_s N_2^\gamma \\ N_1^p \partial_t N_2^\gamma & N_2^p \partial_t N_2^\gamma & N_3^p \partial_t N_2^\gamma & N_4^p \partial_t N_2^\gamma & N_5^p \partial_t N_2^\gamma & N_6^p \partial_t N_2^\gamma & N_7^p \partial_t N_2^\gamma & N_8^p \partial_t N_2^\gamma \\ N_1^p \partial_r N_3^\gamma & N_2^p \partial_r N_3^\gamma & N_3^p \partial_r N_3^\gamma & N_4^p \partial_r N_3^\gamma & N_5^p \partial_r N_3^\gamma & N_6^p \partial_r N_3^\gamma & N_7^p \partial_r N_3^\gamma & N_8^p \partial_r N_3^\gamma \\ N_1^p \partial_s N_3^\gamma & N_2^p \partial_s N_3^\gamma & N_3^p \partial_s N_3^\gamma & N_4^p \partial_s N_3^\gamma & N_5^p \partial_s N_3^\gamma & N_6^p \partial_s N_3^\gamma & N_7^p \partial_s N_3^\gamma & N_8^p \partial_s N_3^\gamma \\ N_1^p \partial_t N_3^\gamma & N_2^p \partial_t N_3^\gamma & N_3^p \partial_t N_3^\gamma & N_4^p \partial_t N_3^\gamma & N_5^p \partial_t N_3^\gamma & N_6^p \partial_t N_3^\gamma & N_7^p \partial_t N_3^\gamma & N_8^p \partial_t N_3^\gamma \\ N_1^p \partial_r N_4^\gamma & N_2^p \partial_r N_4^\gamma & N_3^p \partial_r N_4^\gamma & N_4^p \partial_r N_4^\gamma & N_5^p \partial_r N_4^\gamma & N_6^p \partial_r N_4^\gamma & N_7^p \partial_r N_4^\gamma & N_8^p \partial_r N_4^\gamma \\ N_1^p \partial_s N_4^\gamma & N_2^p \partial_s N_4^\gamma & N_3^p \partial_s N_4^\gamma & N_4^p \partial_s N_4^\gamma & N_5^p \partial_s N_4^\gamma & N_6^p \partial_s N_4^\gamma & N_7^p \partial_s N_4^\gamma & N_8^p \partial_s N_4^\gamma \\ N_1^p \partial_t N_4^\gamma & N_2^p \partial_t N_4^\gamma & N_3^p \partial_t N_4^\gamma & N_4^p \partial_t N_4^\gamma & N_5^p \partial_t N_4^\gamma & N_6^p \partial_t N_4^\gamma & N_7^p \partial_t N_4^\gamma & N_8^p \partial_t N_4^\gamma \\ N_1^p \partial_r N_5^\gamma & N_2^p \partial_r N_5^\gamma & N_3^p \partial_r N_5^\gamma & N_4^p \partial_r N_5^\gamma \\ N_1^p \partial_s N_5^\gamma & N_2^p \partial_s N_5^\gamma & N_3^p \partial_s N_5^\gamma & N_4^p \partial_s N_5^\gamma \\ N_1^p \partial_t N_5^\gamma & N_2^p \partial_t N_5^\gamma & N_3^p \partial_t N_5^\gamma & N_4^p \partial_t N_5^\gamma \\ N_1^p \partial_r N_6^\gamma & N_2^p \partial_r N_6^\gamma & N_3^p \partial_r N_6^\gamma & N_4^p \partial_r N_6^\gamma \\ N_1^p \partial_s N_6^\gamma & N_2^p \partial_s N_6^\gamma & N_3^p \partial_s N_6^\gamma & N_4^p \partial_s N_6^\gamma \\ N_1^p \partial_t N_6^\gamma & N_2^p \partial_t N_6^\gamma & N_3^p \partial_t N_6^\gamma & N_4^p \partial_t N_6^\gamma \\ N_1^p \partial_r N_7^\gamma & N_2^p \partial_r N_7^\gamma & N_3^p \partial_r N_7^\gamma & N_4^p \partial_r N_7^\gamma \\ N_1^p \partial_s N_7^\gamma & N_2^p \partial_s N_7^\gamma & N_3^p \partial_s N_7^\gamma & N_4^p \partial_s N_7^\gamma \\ N_1^p \partial_t N_7^\gamma & N_2^p \partial_t N_7^\gamma & N_3^p \partial_t N_7^\gamma & N_4^p \partial_t N_7^\gamma \\ N_1^p \partial_r N_8^\gamma & N_2^p \partial_r N_8^\gamma & N_3^p \partial_r N_8^\gamma & N_4^p \partial_r N_8^\gamma \\ N_1^p \partial_s N_8^\gamma & N_2^p \partial_s N_8^\gamma & N_3^p \partial_s N_8^\gamma & N_4^p \partial_s N_8^\gamma \\ N_1^p \partial_t N_8^\gamma & N_2^p \partial_t N_8^\gamma & N_3^p \partial_t N_8^\gamma & N_4^p \partial_t N_8^\gamma \\ N_1^p \partial_r N_9^\gamma & N_2^p \partial_r N_9^\gamma & N_3^p \partial_r N_9^\gamma & N_4^p \partial_r N_9^\gamma \\ N_1^p \partial_s N_9^\gamma & N_2^p \partial_s N_9^\gamma & N_3^p \partial_s N_9^\gamma & N_4^p \partial_s N_9^\gamma \\ N_1^p \partial_t N_9^\gamma & N_2^p \partial_t N_9^\gamma & N_3^p \partial_t N_9^\gamma & N_4^p \partial_t N_9^\gamma \end{pmatrix} d\Omega = -\mathbb{G}_{el}^{Q_1 \times Q_1} +$$

$$\int\int\int N_1^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_1 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_1 \partial_r b\ drdsdt \qquad (I.17)$$

$$= \quad \frac{8}{27} - \frac{1}{8}\frac{4}{75} = \frac{1}{1350}(400-9) = \frac{372}{1350} \qquad (I.18)$$

$$\int\int\int N_2^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_2 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_2 \partial_r b\ drdsdt \qquad (I.19)$$

$$= \quad \frac{4}{27} - \frac{1}{8} \qquad (I.20)$$

$$\int\int\int N_3^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_3 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_3 \partial_r b\ drdsdt \qquad (I.21)$$

$$= \quad \frac{2}{27} - \frac{1}{8} \qquad (I.22)$$

$$\int\int\int N_4^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_4 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_4 \partial_r b\ drdsdt \qquad (I.23)$$

$$= \quad \frac{4}{27} \qquad (I.24)$$

$$\int\int\int N_5^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_5 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_5 \partial_r b\ drdsdt \qquad (I.25)$$

$$= \quad \frac{4}{27} \qquad (I.26)$$

$$\int\int\int N_6^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_6 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_6 \partial_r b\ drdsdt \qquad (I.27)$$

$$= \quad \frac{2}{27} \qquad (I.28)$$

$$\int\int\int N_7^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_7 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_7 \partial_r b\ drdsdt \qquad (I.29)$$

$$= \quad \frac{1}{27} \qquad (I.30)$$

$$\int\int\int N_8^p \partial_r N_1^\gamma drdsdt \quad = \quad \int\int\int N_8 \partial_r N_1 drdsdt - \frac{1}{8}\int\int\int N_8 \partial_r b\ drdsdt \qquad (I.31)$$

$$= \quad \frac{2}{27} \qquad (I.32)$$

$$(I.33)$$

For bubble function #1:

$$\mathbb{G}_{el} = \frac{1}{1350} \begin{pmatrix}
372 & 228 & 102 & 198 & 198 & 102 & 43 & 107 \\
372 & 198 & 102 & 228 & 198 & 107 & 43 & 102 \\
372 & 198 & 107 & 198 & 228 & 102 & 43 & 102 \\
-228 & -372 & -198 & -102 & -102 & -198 & -107 & -43 \\
222 & 348 & 252 & 78 & 123 & 182 & 118 & 27 \\
222 & 348 & 182 & 123 & 78 & 252 & 118 & 27 \\
-78 & -222 & -348 & -252 & -27 & -123 & -182 & -118 \\
-78 & -252 & -348 & -222 & -27 & -118 & -182 & -123 \\
147 & 198 & 332 & 198 & 3 & 102 & 268 & 102 \\
222 & 78 & 252 & 348 & 123 & 27 & 118 & 182 \\
-228 & -102 & -198 & -372 & -102 & -43 & -107 & -198 \\
222 & 123 & 182 & 348 & 78 & 27 & 118 & 252 \\
222 & 78 & 27 & 123 & 348 & 252 & 118 & 182 \\
222 & 123 & 27 & 78 & 348 & 182 & 118 & 252 \\
-228 & -102 & -43 & -102 & -372 & -198 & -107 & -198 \\
-78 & -222 & -123 & -27 & -252 & -348 & -182 & -118 \\
147 & 198 & 102 & 3 & 198 & 332 & 268 & 102 \\
-78 & -252 & -118 & -27 & -222 & -348 & -182 & -123 \\
-3 & -147 & -198 & -102 & -102 & -198 & -332 & -268 \\
-3 & -102 & -198 & -147 & -102 & -268 & -332 & -198 \\
-3 & -102 & -268 & -102 & -147 & -198 & -332 & -198 \\
147 & 3 & 102 & 198 & 198 & 102 & 268 & 332 \\
-78 & -27 & -123 & -222 & -252 & -118 & -182 & -348 \\
-78 & -27 & -118 & -252 & -222 & -123 & -182 & -348 \\
-576 & 576 & 384 & -384 & -384 & 384 & 256 & -256 \\
-576 & -384 & 384 & 576 & -384 & -256 & 256 & 384 \\
-576 & -384 & -256 & -384 & 576 & 384 & 256 & 384
\end{pmatrix}$$

Considering a 2x2x2 macroelement of size 4x4x4. Then NV=3*3*3+8=35, NP=3*3*3=27 Matrix $\mathbb{G}$ is 3*35x27=105*27

After bc are imposed on all nodes on the boundary, 9 Vnodes are still free (8 bubble nodes and the node in the middle), i.e. 9*3 dofs = 27 and there are 3x3x3=27 pressure nodes. So $\tilde{G}$ is 27*27. no less.

We get

$$\tilde{G} = \frac{1}{1350} \left(\ jfgljhhgl \ \right)$$

I have tried all kinds of bubbles but I usually arrive at a null space of dimension 3 to 5... typically for bubble 1, dim=5, while for bubble=2 dim=3.

Appendix J

Computational Geophysics GEO4-1427 - projects

J.0.1 Convection in a box *

This exercise builds on your existing 2D advection-diffusion code. Scale up the benchmark described in Section 19.21 so that it runs in a 1000x1000 km domain with Earth-like parameters and velocities (the maximum velocity is denoted by ν_{conv} and will be varied). Start with an initial zero temperature field and Earth-like boundary conditions on the top and bottom, e.g. $T = 20$ at the top and $T = 1000$ at the bottom. Set $k = 3$, $C_p = 1250$ and $\rho = 3000$.

Run the code until steady state is reached. Implement an algorithm which computes the average temperature

$$< T >= \frac{1}{L_x L_y} \iint T(x,y) dx dy$$

in the domain and plot it as a function of time. Also compute the root mean square velocity in the domain:

$$\nu_{rms} = \sqrt{\frac{1}{L_x L_y} \iint (u^2 + v^2) dx dy}$$

Plot the steady state $< T >$ and ν_{rms} as a function of the resolution h. Plot the temperature on the $x = L_x/2$ line for different values of ν_{conv}. When possible, make a link with the Mantle Dynamics practical.

Bonus: Compute and plot the heat flux $\vec{q} = -\vec{\nabla} T$ in the center of the elements.

J.0.2 Corner flow subduction *

- In this experiment the velocity is prescribed in geometrically simple subducting and overriding plates, while the velocity in the mantle is computed by means of an analytical solution coined 'corner flow velocity'. Details are to be found in G.K. Batchelor, *An introduction to fluid dynamics*. Cambridge University Press, 1967.

- Write a function which prescribes the velocity in a lithospheric sized domain.

- Use this velocity to drive the system in time (choose the appropriate values for the coefficients in the heat transport equation)

- prescribe a constant temperature value at the top, and fix the temperature on both sides, but only in the plates (along lengths l_1 and l_2. Choose an appropriate plate temperature model.

- Run the model over millions of years with different velocities.

- Measure the depth of the isotherm 800° as a function of time (bonus).

- Is steady state ever reached ?

If all goes well, you should be able to recover similar results:

3D setup with prescribed velocity. b) example of temperature field evolution
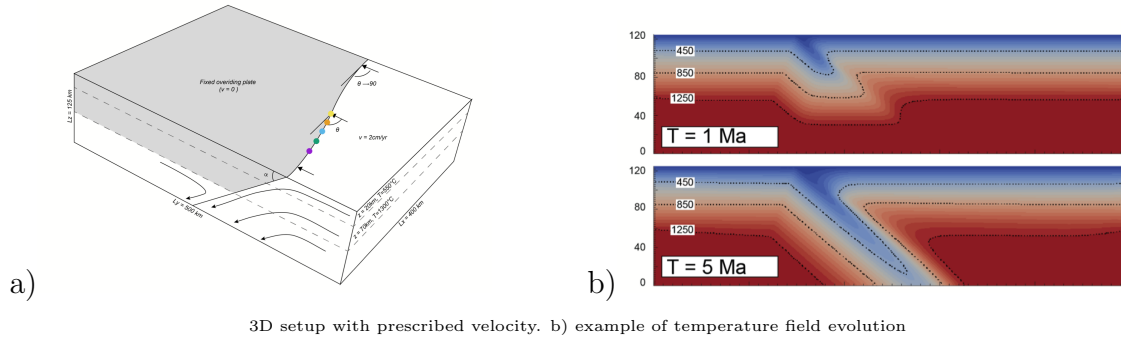
what are l1, l2 ? rephrase !

J.0.3   From 2D to 3D **

Rewrite your 2D FEM advection-diffusion code so that it now runs on a cube. You will need to create a new connectivity array, compute new elemental matrices, etc ... Center the cube on the origin of the axis system.

Compute the steady-state solution of a problem without heat advection. The domain is a unit cube, $k = \rho = C_p = 1$. A temperature $T_{max} = 1$ is prescribed at the bottom and $T_{min} = 0$ at the top.

Same problem when these boundary conditions are now prescribed on the faces.

Prescribe an temperature field such that it is 1 everywhere in the domain but 2 inside a sphere of radius 0.1 centered at $(0.66, 0.66, 0.66)$. This time no diffusion takes place but we wish to advect the field using the velocity $\vec{v} = (y, x, 0)$. What is the highest resolution that is achievable on your computer?

J.0.4   Triangular linear elements */**

Redo the 2D advection-diffusion exercises with triangular elements. You will need to make a new icon array, and recompute the mass matrix and other matrices. The triangular elements are constructed by splitting square elements along the diagonal. See Section 8.3.10 for the basis functions and their derivatives. See Appendix C.0.5 for the calculations of the matrices.

J.0.5   Triangular linear elements ***

Same exercise as above, with an additional task: run the benchmark presented in Section 20.13. For this you will need to generate a mesh such that nodes are placed on the perimeter of the cylinder and there is no node inside the cylinder:
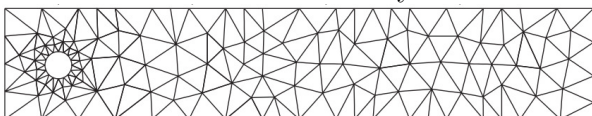
You can build it by hand, or you can use an external mesher library (see Delaunay triangulation inside scipy). Vary the heat conduction coefficient to show the effect of diffusion on the obtained steady state temperature field.

J.0.6 Diffusion of topography ****

In a 2D plane assign each node an initial topography $h(x, y, t = 0)$ given by

$$h(x, y, t = 0) = h_0 \sin(\pi x / L_x) + \xi(x, y) \delta h$$

where L_x and L_y are the dimensions of the domain, h_0 is the height of the orogen, $xi(x, y)$ is a random perturbation in $[-1, 1]$ and δh is the amplitude of the perturbation.

We wish to 'erode' the topography by means of a (nonlinear) diffusion law as in section 2.1.1 of Burov & Cloetingh [197].

1. What are the physical parameters needed to carry out this experiment? What are the appropriate boundary conditions? What is the steady state? What are the relevant time scales? How should we choose the time step?

2. Write a code which solves the linear diffusion equation until steady state is reached. Explore the effect of δh. Compute the slope $\vec{\nabla} h$ inside each element and plot its time evolution.

3. Implement the nonlinear diffusion law and run the model once again.

4. If a source term is added to the diffusion equation it is in fact a vertical velocity ($\partial h / \partial t$ has the dimensions of a velocity). Add a source term which generates uplift in a symmetric and asymmetric manner.

Relevant Literature[1348] [1293], also check Appendix H.

J.0.7 An example of a hand-built triangular mesh

We start from a 8x5 domain which is tesselated as follows:
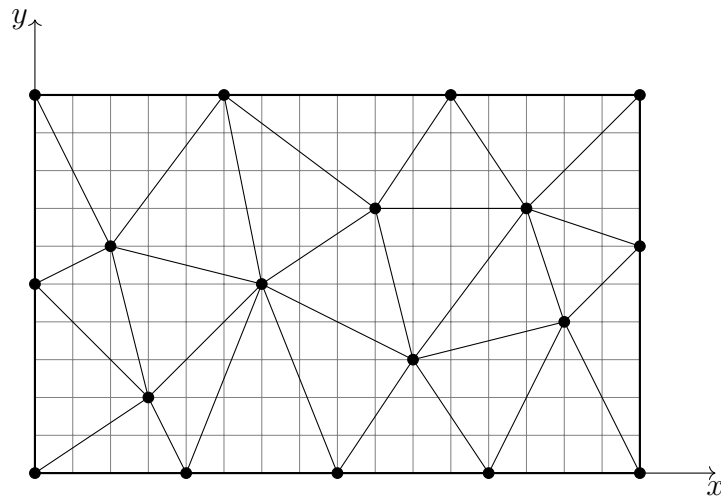
We can first label the nodes:
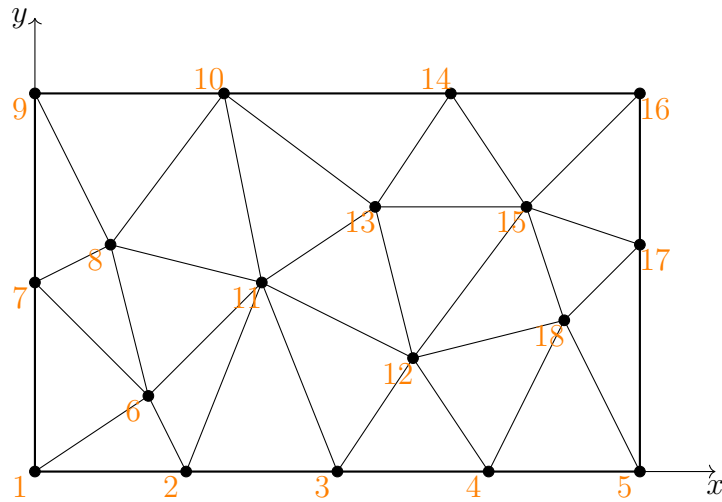
and then label the elements:
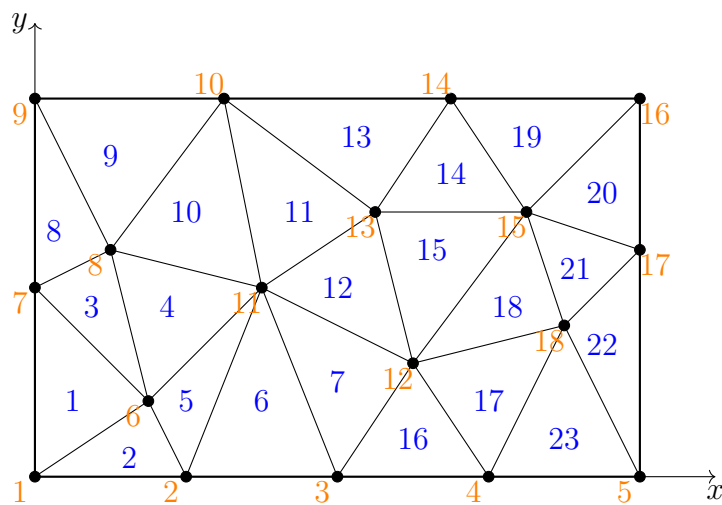
We can finally build the connectivity array by hand:

icon(1,1)=1
icon(2,1)=6
icon(3,1)=7
icon(1,2)=1
icon(2,2)=2
icon(3,2)=6
icon(1,3)=7
icon(2,3)=6
icon(3,3)=8

...

icon(1,12)=11
icon(2,12)=12
icon(3,12)=13

...

icon(1,19)=14
icon(2,19)=15
icon(3,19)=16

The labelling of nodes and elements above is done by a human so it starts at 1. When implementing this in python, you know what to do ...

J.0.8 How to visualise data on a triangular mesh with Paraview

If arrays `x,y` contain the coordinates of the nodes, your connectivity array is called `icon`, and your mesh consistes of `nel` elements and comprises `nnp` nodes, you can use the following code to generate a vtu file to be opened with Paraview. You also need a temperature array `T`.

Code at `https: //raw.githubusercontent.com/cedrict/fieldstone/master/images/compgeo/mesh_visu.py`

```python
vtufile=open('mesh.vtu',"w")
vtufile.write("<VTKFile type='UnstructuredGrid' version='0.1' byte_order='BigEndian'> \n")
vtufile.write("<UnstructuredGrid> \n")
vtufile.write("<Piece NumberOfPoints=' %5d ' NumberOfCells=' %5d '> \n" %(nnp,nel))
vtufile.write("<Points> \n")
vtufile.write("<DataArray type='Float32' NumberOfComponents='3' Format='ascii'> \n")
for i in range(0,nnp):
    vtufile.write("%10e %10e %10e \n" %(x[i],y[i],0.))
vtufile.write("</DataArray>\n")
vtufile.write("</Points> \n")
#vtufile.write("<CellData Scalars='scalars'>\n")
#vtufile.write("<DataArray type='Float32' Name='area' Format='ascii'> \n")
#for iel in range (0,nel):
#    vtufile.write("%10e\n" % (area[iel]))
#vtufile.write("</DataArray>\n")
#vtufile.write("</CellData>\n")
vtufile.write("<PointData Scalars='scalars'>\n")
vtufile.write("<DataArray type='Float32' Name='T' Format='ascii'> \n")
for i in range(0,nnp):
    vtufile.write("%10e \n" %T[i])
vtufile.write("</DataArray>\n")
vtufile.write("</PointData>\n")
vtufile.write("<Cells>\n")
vtufile.write("<DataArray type='Int32' Name='connectivity' Format='ascii'> \n")
for iel in range (0,nel):
    vtufile.write("%d %d %d\n" %(icon[0,iel],icon[1,iel],icon[2,iel]))
vtufile.write("</DataArray>\n")
vtufile.write("<DataArray type='Int32' Name='offsets' Format='ascii'> \n")
for iel in range (0,nel):
    vtufile.write("%d \n" %((iel+1)*3))
vtufile.write("</DataArray>\n")
vtufile.write("<DataArray type='Int32' Name='types' Format='ascii'>\n")
for iel in range (0,nel):
    vtufile.write("%d \n" %5)
vtufile.write("</DataArray>\n")
vtufile.write("</Cells>\n")
vtufile.write("</Piece>\n")
vtufile.write("</UnstructuredGrid>\n")
vtufile.write("</VTKFile>\n")
vtufile.close()
```

Appendix K

Using prisms in forward gravity modelling

This appendix was written by Sverre Hassing as part of his Bachelor thesis. Although the final formula are definitely correct, the derivations below may still contain a typo.

The derivations for prisms have been published in the early 50's [868]. However, to the best of our knowledge the full derivation has not been carried out in English in full detail. The derivations are based on those of Mader (1951) [868] and of Nagy *et al.* (2000) [976, 975]. Mader provided the derivations in some detail, while Nagy *et al.* interpreted the results in a more modern style.

K.0.1 Basic formulas

The derivations for prisms are a lot more complicated than that for the point masses. We start with the following two integral equations which are integral part of the derivations:

$$
\int \frac{x^2 dx}{x^2 + z^2} = x - z \arctan \frac{x}{z} \tag{K.1}
$$

$$
\int \frac{dx}{\sqrt{x^2 + y^2 + z^2}} = \ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) \tag{K.2}
$$

Another equation that will come back multiple times is of the form:

$$
\int \frac{du}{(v^2 + w^2)\sqrt{u^2 + v^2 + w^2}} \tag{K.3}
$$

This can be solved with a trigonometric substitution, where $u = \sqrt{v^2 + w^2} \tan\phi$. This means that $du = \frac{\sqrt{v^2+w^2}}{cos^2\phi} d\phi$.

$$
\begin{aligned}
\int \frac{du}{(v^2 + w^2)\sqrt{u^2 + v^2 + w^2}} &= \int \frac{\sqrt{v^2 + w^2}}{cos^2\phi} \frac{1}{(v^2 + w^2)\tan^2\phi + v^2} \frac{d\phi}{\sqrt{v^2 + w^2 + (v^2 + w^2)\tan^2\phi}} \\
&= \int \frac{\sqrt{v^2 + w^2}}{cos\phi^2} \frac{1}{v^2(\tan^2\phi + 1) + w^2\tan^2\phi} \frac{d\phi}{\sqrt{(v^2 + w^2)(\tan^2\phi + 1)}} \\
&= \int \frac{\sqrt{v^2 + w^2}}{cos\phi^2} \frac{1}{\frac{v^2 + w^2\sin^2\phi}{cos^2\phi}} \frac{d\phi}{\frac{\sqrt{v^2+w^2}}{cos\phi}} \\
&= \int \frac{\cos\phi \, d\phi}{v^2 + w^2\sin^2\phi} \tag{K.4}
\end{aligned}
$$

1338

A second substitution is needed where $t = \frac{w}{v} \sin \phi$ and $dt = \frac{v}{w} \cos \phi d\phi$:

$$
\int \frac{\cos \phi d\phi}{v^2 + w^2 \sin^2 \phi} = \int \frac{vdt}{w(v^2 + v^2 t^2)}
$$

$$
= \frac{1}{vw} \int \frac{dt}{1 + t^2}
$$

$$
= \frac{1}{vw} \arctan t
$$

$$
= \frac{1}{vw} \arctan \frac{w \sin \phi}{v} \tag{K.5}
$$

Now the $\sin \phi$ needs to be converted back to u, v, w. If it is known that $\tan \phi = \frac{u}{\sqrt{v^2 + w^2}}$, then it follows that $\sin \phi = \frac{u}{\sqrt{u^2 + v^2 + w^2}}$. Eq.(K.5) then becomes

$$
\frac{1}{vw} \arctan \frac{w \sin \phi}{v} = \frac{1}{vw} \arctan \frac{uw}{v\sqrt{u^2 + v^2 + w^2}} \tag{K.6}
$$

and finally

$$
\boxed{\int \frac{du}{(v^2 + w^2)\sqrt{u^2 + v^2 + w^2}} = \frac{1}{vw} \arctan \frac{uw}{v\sqrt{u^2 + v^2 + w^2}}} \tag{K.7}
$$

K.0.2 The gravitational potential

Each prism is assumed to have constant density ρ. The gravitational potential is integrated over the whole volume of the prism:

$$
U(P) = -\mathcal{G}\rho \underbrace{\int_{x_1}^{x_2} \int_{y_1}^{y_2} \int_{z_1}^{z_2} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}}}_{I} \tag{K.8}
$$

In what follows we work out the exact form for the triple integral term. Elementary Eq. (K.2) can be applied to the integral for dx in Eq. (K.8).

$$
I = \iiint \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}}
$$

$$
= \iint \left(\int \frac{dx}{\sqrt{x^2 + y^2 + z^2}} \right) dydz
$$

$$
= \iint \ln \left(x + \sqrt{x^2 + y^2 + z^2} \right) dydz \tag{K.9}
$$

We further proceed with the integration with respect to y. We define

$$
\frac{f = \int \ln \left(x + \sqrt{x^2 + y^2 + z^2} \right) dz}{f' = \frac{y}{\left(x + \sqrt{x^2 + y^2 + z^2} \right)\sqrt{x^2 + y^2 + z^2}}} \;\Bigg|\; \frac{g' = dy}{g = y}
$$

and using $\int fg' = fg - \int fg'$ we have

$$
I = \underbrace{y \int \ln \left(x + \sqrt{x^2 + y^2 + z^2} \right) dz}_{A} - \underbrace{\iint \frac{y^2 dz}{\left(x + \sqrt{x^2 + y^2 + z^2} \right) \sqrt{x^2 + y^2 + z^2}} dy}_{B} \tag{K.10}
$$

The calculation of I is then split into two large integrals denoted A and B, calculated in the following subsections. Note that we have not made use of the integral bounds yet.

1339

The calculation of A

The first step in calculating A is to carry out a similar partial integration as seen before.

$$
\begin{array}{c|c}
f = \ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) & g' = dz \\
\hline
f' = \frac{\partial f}{\partial z} = \frac{z}{\left(x + \sqrt{x^2 + y^2 + z^2}\right)\sqrt{x^2 + y^2 + z^2}} & g = z
\end{array}
$$

$$
\begin{aligned}
A &= y\left(z\ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) - \int \frac{z^2 dz}{\left(x + \sqrt{x^2 + y^2 + z^2}\right)\sqrt{x^2 + y^2 + z^2}}\right) \\
&= \underbrace{yz\ln\left(x + \sqrt{x^2 + y^2 + z^2}\right)}_{A_0} - \underbrace{y\int \frac{z^2 dz}{\left(x + \sqrt{x^2 + y^2 + z^2}\right)\sqrt{x^2 + y^2 + z^2}}}_{A_1} \tag{K.11}
\end{aligned}
$$

We now focus on the A_1 integral. We first multiply the numerator and denominator by $-x + \sqrt{x^2 + y^2 + z^2}$. The last step uses Eqs. (K.2), (K.7) and (K.1) respectively for each term.

$$
\begin{aligned}
A_1 &= \int \frac{z^2 dz}{\left(x + \sqrt{x^2 + y^2 + z^2}\right)\sqrt{x^2 + y^2 + z^2}} \frac{-x + \sqrt{x^2 + y^2 + z^2}}{-x + \sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{(-xz^2 + z^2\sqrt{x^2 + y^2 + z^2})dz}{(x^2 + y^2 + z^2 - x^2)\sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{-xz^2 dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} + \int \frac{z^2\sqrt{x^2 + y^2 + z^2}dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{-x\left(z^2 + y^2 - y^2\right)dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} + \int \frac{z^2 dz}{y^2 + z^2} \\
&= \int \frac{-x dz}{\sqrt{x^2 + y^2 + z^2}} + \int \frac{xy^2 dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} + \int \frac{z^2 dz}{y^2 + z^2} \\
&= -x\int \frac{dz}{\sqrt{x^2 + y^2 + z^2}} + xy^2\int \frac{dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} + \int \frac{z^2 dz}{y^2 + z^2} \\
&= -x\ln\left(z + \sqrt{x^2 + y^2 + z^2}\right) + y\arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} + z - y\arctan\frac{z}{y} \tag{K.12}
\end{aligned}
$$

This can be combined to get the final expression for A:

$$
A = y\left(z\ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) + x\ln\left(z + \sqrt{x^2 + y^2 + z^2}\right) - y\arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} - z + y\arctan\frac{z}{y}\right) \tag{K.13}
$$

The last two terms can be left out because they will cancel out when computing the integration boundaries from x_1 to x_2, because these terms do not contain the variable x. Finally we arrive at the following expression for A:

$$
A = yz\ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) + xy\ln\left(z + \sqrt{x^2 + y^2 + z^2}\right) - y^2\arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} \tag{K.14}
$$

The calculation of B

The inner integral can be simplified similarly to how A_1 was simplified in Eq. (K.12), by multiplying both numerator and denominator with $-x + \sqrt{x^2 + y^2 + z^2}$. The last step uses Eqs. (K.7) and (K.1):

$$
\begin{aligned}
B &= -\int y^2 \int \frac{dz}{\left(x + \sqrt{x^2 + y^2 + z^2}\right)\sqrt{x^2 + y^2 + z^2}} \frac{-x + \sqrt{x^2 + y^2 + z^2}}{-x + \sqrt{x^2 + y^2 + z^2}} dy \\
&= -\int y^2 \int \frac{-x + \sqrt{x^2 + y^2 + z^2}}{(x^2 + y^2 + z^2 - x^2)\sqrt{x^2 + y^2 + z^2}} dz\, dy \\
&= -\int y^2 \left(-\int \frac{x\, dz}{(y^2 + z^2)\sqrt{x^2 + y^2 + z^2}} + \int \frac{dz}{y^2 + z^2}\right) dy \\
&= -\int y^2 \left(-\frac{1}{y}\arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} + \frac{1}{y}\arctan\frac{z}{y}\right) dy \\
&= \int y \arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} dy
\end{aligned}
$$

(K.15)

Again the second term can be left out, because it does not contain the variable x. The next step is to apply a partial integration to B.

$$
\begin{array}{c|c}
f = \arctan\dfrac{xz}{y\sqrt{x^2+y^2+z^2}} & g' = y \\[2ex]
\hline
f' = -xz\dfrac{\frac{1}{y^2\sqrt{x^2+y^2+z^2}} + \frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}}}{\frac{x^2 z^2}{y^2(x^2+y^2+z^2)} + 1} & g = \dfrac{y^2}{2}
\end{array}
$$

$$
B = \frac{y^2}{2}\arctan\frac{xz}{y\sqrt{x^2 + y^2 + z^2}} + \frac{xz}{2}\underbrace{\int y^2 \frac{\frac{1}{y^2\sqrt{x^2+y^2+z^2}} + \frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}}}{\frac{x^2 z^2}{y^2(x^2+y^2+z^2)} + 1}}_{B_1}
$$

(K.16)

Let us finish by calculating the integral B_1:

$$
\begin{aligned}
B_1 &= \frac{xz}{2} \int y^2 \frac{\frac{1}{y^2\sqrt{x^2+y^2+z^2}} + \frac{1}{(x^2+y^2+z^2)^{3/2}}}{\frac{x^2 z^2}{y^2(x^2+y^2+z^2)} + 1} dy \\
&= \frac{xz}{2} \int y^2 \frac{\frac{x^2+y^2+z^2}{y^2(x^2+y^2+z^2)^{3/2}} + \frac{y^2}{y^2(x^2+y^2+z^2)^{3/2}}}{\frac{x^2 z^2 + y^2(x^2+y^2+z^2)}{y^2(x^2+y^2+z^2)}} dy \\
&= \frac{xz}{2} \int y^2 \frac{\frac{x^2+2y^2+z^2}{y^2(x^2+y^2+z^2)^{3/2}}}{\frac{x^2 z^2 + y^2(x^2+y^2+z^2)}{y^2(x^2+y^2+z^2)}} dy \\
&= \frac{xz}{2} \int y^2 \frac{x^2+2y^2+z^2}{\sqrt{x^2+y^2+z^2}(x^2 z^2 + y^2(x^2+y^2+z^2))} dy \\
&= \frac{xz}{2} \int y^2 \frac{x^2+2y^2+z^2}{\sqrt{x^2+y^2+z^2}(x^2+y^2)(z^2+y^2)} dy \\
&= \frac{xz}{2} \left(\int \frac{2dy}{\sqrt{x^2+y^2+z^2}} + \int \frac{-(x^2+z^2)y^2 - 2x^2 z^2}{(x^2+y^2)(y^2+z^2)\sqrt{x^2+y^2+z^2}} dy \right) \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{xz}{2} \int \frac{(x^2+z^2)y^2 + 2x^2 z^2}{(x^2+y^2)(y^2+z^2)\sqrt{x^2+y^2+z^2}} dy \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{xz}{2} \int \frac{x^2 y^2 + y^2 z^2 + 2x^2 z^2}{(x^2+y^2)(y^2+z^2)\sqrt{x^2+y^2+z^2}} dy \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{xz}{2} \int \frac{x^2(y^2+z^2) + z^2(x^2+y^2)}{(x^2+y^2)(y^2+z^2)\sqrt{x^2+y^2+z^2}} dy \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{xz}{2} \int \frac{x^2}{(x^2+y^2)\sqrt{x^2+y^2+z^2}} dy - \frac{xz}{2} \int \frac{z^2}{(y^2+z^2)\sqrt{x^2+y^2+z^2}} dy \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{xz}{2} \frac{x^2 \arctan \frac{yz}{x\sqrt{x^2+y^2+z^2}}}{xz} - \frac{xz}{2} \frac{z^2 \arctan \frac{xy}{z\sqrt{x^2+y^2+z^2}}}{xz} \\
&= xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) - \frac{x^2}{2} \arctan \frac{yz}{x\sqrt{x^2+y^2+z^2}} - \frac{z^2}{2} \arctan \frac{xy}{z\sqrt{x^2+y^2+z^2}} \qquad \text{(K.17)}
\end{aligned}
$$

This can be combined to get the full expression for B:

$$
B = xz \ln\left(\sqrt{x^2+y^2+z^2} + y \right) - \frac{x^2}{2} \arctan \frac{zy}{x\sqrt{x^2+y^2+z^2}} + \frac{y^2}{2} \arctan \frac{xz}{y\sqrt{x^2+y^2+z^2}} - \frac{z^2}{2} \arctan \frac{xy}{x\sqrt{z^2+y}}
$$

Combining A and B

Now A and B can be combined to get the expression of I

$$
\begin{aligned}
I &= A + B \\
&= yz \ln\left(x + \sqrt{x^2+y^2+z^2} \right) + xy \ln\left(z + \sqrt{x^2+y^2+z^2} \right) - y^2 \arctan \frac{xz}{y\sqrt{x^2+y^2+z^2}} \\
&+ xz \ln\left(\sqrt{x^2+y^2+z^2} + y \right) - \frac{x^2}{2} \arctan \frac{zy}{x\sqrt{x^2+y^2+z^2}} + \frac{y^2}{2} \arctan \frac{xz}{y\sqrt{x^2+y^2+z^2}} - \frac{z^2}{2} \arctan \frac{}{x\sqrt{z}} \\
&= yz \ln\left(x + \sqrt{x^2+y^2+z^2} \right) + xy \ln\left(z + \sqrt{x^2+y^2+z^2} \right) + xz \ln\left(y + \sqrt{x^2+y^2+z^2} \right) \\
&- \frac{x^2}{2} \arctan \frac{zy}{x\sqrt{x^2+y^2+z^2}} - \frac{y^2}{2} \arctan \frac{xz}{y\sqrt{x^2+y^2+z^2}} - \frac{z^2}{2} \arctan \frac{xy}{x\sqrt{z^2+y^2+z^2}}
\end{aligned}
$$

1342

The boundaries for the volume from Eq. (**??**) need to be applied to the result of the integration. The boundary conditions are computed by plugging the upper value into the equation and subtracting the equation with the lower value plugged in. When the upper and lower values are respectively x_2 and x_1 for some function $f(x)$, this is $f(x_2) - f(x_1)$. This can be represented more efficiently with a summation over the subscript. Something needs to be added to still keep the subtraction in there. This can be done by adding a factor of -1^i, where i is the summation index. This will be positive when i is even and negative when i is odd. The new way of showing the result would be $\sum_{i=1}^{2} -1^i f(x_i)$. This is especially useful when there are three different integration boundaries to resolve. r will be used instead of $\sqrt{x^2 + y^2 + z^2}$.

$$
\begin{aligned}
I &= \left|\left|\left| yz \ln(x+r) + xy \ln(z+r) + xz \ln(y+r) - \frac{x^2}{2}\arctan\frac{zy}{xr} - \frac{y^2}{2}\arctan\frac{xz}{yr} - \frac{z^2}{2}\arctan\frac{xy}{xr} \right|_{x_1}^{x_2}\right|_{y_1}^{y_2}\right|_{z_1}^{z_2} \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(y_j z_k \ln(x_i + r_{ijk}) + x_i y_j \ln(z_k + r_{ijk}) + x_i z_k \ln(y_j + r_{ijk}) \right. \\
&\qquad \left. - \frac{x_i^2}{2}\arctan\frac{z_k y_j}{x_i r_{ijk}} - \frac{y_j^2}{2}\arctan\frac{x_i z_k}{y_j r_{ijk}} - \frac{z_k^2}{2}\arctan\frac{x_i y_j}{x_i r_{ijk}} \right)
\end{aligned}
\tag{K.19}
$$

> There is probably a mistake in eq above and below, last term, most likely should contain zk in denominator?

Finally,

$$
\begin{aligned}
U(\vec{r}) &= \mathcal{G}\rho \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(y_j z_k \ln(x_i + r_{ijk}) + x_i y_j \ln(z_k + r_{ijk}) + x_i z_k \ln(y_j + r_{ijk}) \right. \\
&\qquad \left. - \frac{x_i^2}{2}\arctan\frac{z_k y_j}{x_i r_{ijk}} - \frac{y_j^2}{2}\arctan\frac{x_i z_k}{y_j r_{ijk}} - \frac{z_k^2}{2}\arctan\frac{x_i y_j}{x_i r_{ijk}} \right)
\end{aligned}
\tag{K.20}
$$

K.0.3 The gravity vector $\vec{g}$

In 3D Cartesian coordinates the gravity vector is expressed as

$$
\vec{g} = -\vec{\nabla}U = \begin{pmatrix} -\frac{\partial U}{\partial x} \\ -\frac{\partial U}{\partial y} \\ -\frac{\partial U}{\partial z} \end{pmatrix}
\tag{K.21}
$$

The easiest way to calculate this is by including the partial derivatives in the original integral (K.8).

$$
\begin{aligned}
I_x(\vec{r}) &= \iiint \frac{\partial}{\partial x}\frac{dxdydz}{\sqrt{x^2+y^2+z^2}} \\
&= -\iiint \frac{x\,dxdydz}{(\sqrt{x^2+y^2+z^2})^3} \\
&= \iint \frac{dydz}{\sqrt{x^2+y^2+z^2}}
\end{aligned}
\tag{K.22}
$$

The integral (K.2) can be used, followed by the calculation of A as seen in Section K.0.2 without the multiplication with y:

$$
\begin{aligned}
I_x(\vec{r}) &= \int \ln\left(x + \sqrt{x^2 + y^2 + z^2}\right) dz \\
&= z \ln\left(y + \sqrt{x^2 + y^2 + z^2}\right) + y \ln\left(z + \sqrt{x^2 + y^2 + z^2}\right) - x \arctan \frac{yz}{x\sqrt{x^2 + y^2 + z^2}} \quad \text{(K.23)}
\end{aligned}
$$

The integration boundaries can be applied. Multiplication with $\mathcal{G}$ and ρ is the final step in deriving the element of the gravity vector component (g_x).

$$
g_x = \mathcal{G}\rho \sum_{i,j,k=1}^{2} (-1)^{i+j+k} \left(z_k \ln\left(y_j + \sqrt{x_i^2 + y_j^2 + z_k^2}\right) + y_j \ln\left(z_j + \sqrt{x_i^2 + y_j^2 + z_k^2}\right) - x_i \arctan \frac{y_j z_k}{x_i\sqrt{x_i^2 + y_j^2 + }} \right.
$$

The same can be done for the $y-$ and $z-$components and in the end we obtain

$$
\begin{aligned}
g_x &= \mathcal{G}\rho \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(z_k \ln(y_j + r_{ijk}) + y_j \ln(z_j + r_{ijk}) - x_i \arctan\frac{y_j z_k}{x_i r_{ijk}}\right) \\
g_y &= \mathcal{G}\rho \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(z_k \ln(x_i + r_{ijk}) + x_i \ln(z_j + r_{ijk}) - y_j \arctan\frac{x_i z_k}{y_j r_{ijk}}\right) \\
g_z &= \mathcal{G}\rho \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(x_i \ln(y_j + r_{ijk}) + y_j \ln(x_i + r_{ijk}) - z_k \arctan\frac{x_i y_j}{z_k r_{ijk}}\right)
\end{aligned}
$$

These equations can be found in various other papers such as Eq. (6) in Heck and Seitz, 2007 [590], Eqs. (8,11,12) in Nagy *et al.* , 2000 [976] (note that there is a mistake there that is later fixed in [975]), appendix A in Couder-Castaneda *et al.* ., 2015 [298] and the derivation between (14) and (15) in Mader, 1951 [868].

K.0.4 The gravity gradient tensor

The different elements of the gravity gradient tensor can be determined by partially differentiating each component of the gravity vector with respect to each space coordinate. As will be shown later, $\boldsymbol{T}$ should be a symmetric matrix and its trace should equal zero.

The diagonal terms

$$
T_{xx} = \frac{\partial}{\partial x}g_x = -\frac{\partial^2}{\partial x^2}U(\vec{r}) = \mathcal{G}\rho\frac{\partial^2}{\partial x^2}\left(-I(\vec{r})\right) \quad \text{(K.24)}
$$

$$
\begin{aligned}
I_{xx}(\vec{r}) &= \iiint \frac{\partial^2}{\partial x^2}\frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \iint \frac{\partial}{\partial x}\frac{dydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= -\iint \frac{xdydz}{\sqrt{x^2 + y^2 + z^2}^3} \quad \text{(K.25)}
\end{aligned}
$$

A trigonometric substitution is applied to solve this integral. This uses $y = \sqrt{x^2 + z^2}\tan\phi$ and $dy = \frac{\sqrt{x^2+y^2}}{\cos^2\phi}d\phi$.

$$
\begin{aligned}
I_{xx} &= -x\iint \frac{\sqrt{x^2+y^2}}{\cos^2\phi}\frac{d\phi dz}{\sqrt{(x^2+z^2)\tan^2\phi + x^2 + z^2}^3} \\
&= -x\iint \frac{\sqrt{x^2+y^2}}{\cos^2\phi}\frac{d\phi dz}{\sqrt{(x^2+z^2)(\tan^2\phi + 1)}^3} \\
&= -x\iint \frac{\sqrt{x^2+y^2}}{\cos^2\phi}\frac{d\phi dz}{\left(\frac{\sqrt{(x^2+z^2)}}{\cos\phi}\right)^3} \\
&= -x\int \frac{1}{x^2+z^2}\int \cos\phi\, d\phi dz \\
&= -x\int \frac{1}{x^2+z^2}\sin\phi dz
\end{aligned}
\tag{K.26}
$$

Now the substitution needs to be undone. If $\tan\phi = \frac{y}{\sqrt{x^2+z^2}}$, then $\sin\phi = \frac{y}{\sqrt{x^2+y^2+z^2}}$ and then

$$
I_{xx} = -xy\int \frac{dz}{(x^2+z^2)\sqrt{x^2+y^2+z^2}}
\tag{K.27}
$$

This can be solved by applying equation (K.7).

$$
\begin{aligned}
I_{xx} &= -\frac{xy}{xy}\arctan\frac{yz}{x\sqrt{x^2+y^2+z^2}} \\
&= -\arctan\frac{yz}{x\sqrt{x^2+y^2+z^2}}
\end{aligned}
$$

The tensor element T_{xx} is then formulated as follows (the other elements of the diagonal are found by cyclic permutation of x, y and z):

$$
\begin{aligned}
T_{xx} &= \mathcal{G}\rho\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(-\arctan\frac{y_j z_k}{x_i r_{ijk}}\right) \\
T_{yy} &= \mathcal{G}\rho\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(-\arctan\frac{x_i z_k}{y_j r_{ijk}}\right) \\
T_{zz} &= \mathcal{G}\rho\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(-1)^{i+j+k}\left(-\arctan\frac{x_i y_j}{z_k r_{ijk}}\right)
\end{aligned}
$$

The off-diagonal terms of the tensor

The other elements are easier to calculate, because the partial derivatives cancel out the integrals:

$$
\begin{aligned}
I_{xy} &= \iiint \frac{\partial^2}{\partial x \partial y} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \iint \frac{\partial}{\partial y} \frac{dydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{dz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \ln\left(z + \sqrt{x^2 + y^2 + z^2}\right)
\end{aligned}
\tag{K.28}
$$

$$
\begin{aligned}
I_{xz} &= \iiint \frac{\partial^2}{\partial x \partial z} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \iint \frac{\partial}{\partial z} \frac{dydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{dy}{\sqrt{x^2 + y^2 + z^2}} \\
&= \ln\left(y + \sqrt{x^2 + y^2 + z^2}\right)
\end{aligned}
\tag{K.29}
$$

$$
\begin{aligned}
I_{yz} &= \iiint \frac{\partial^2}{\partial y \partial z} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \iint \frac{\partial}{\partial z} \frac{dxdz}{\sqrt{x^2 + y^2 + z^2}} \\
&= \int \frac{dx}{\sqrt{x^2 + y^2 + z^2}} \\
&= \ln\left(x + \sqrt{x^2 + y^2 + z^2}\right)
\end{aligned}
\tag{K.30}
$$

From these calculations it should be obvious why $\boldsymbol{T}$ is a symmetric tensor. When applying the second partial derivatives, their order does not matter:

$$
I_{xy} = \iiint \frac{\partial^2}{\partial x \partial y} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} = \iiint \frac{\partial^2}{\partial y \partial x} \frac{dxdydz}{\sqrt{x^2 + y^2 + z^2}} = I_{yx}
\tag{K.31}
$$

The tensor elements following from this are:

$$
\begin{aligned}
T_{xy} = T_{yx} &= \mathcal{G}\rho \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} (-1)^{i+j+k} \left(\ln\left(z_k + r_{ijk}\right)\right) \\
T_{xz} = T_{zx} &= \mathcal{G}\rho \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} (-1)^{i+j+k} \left(\ln\left(y_j + r_{ijk}\right)\right) \\
T_{yz} = T_{zy} &= \mathcal{G}\rho \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} (-1)^{i+j+k} \left(\ln\left(x_i + r_{ijk}\right)\right)
\end{aligned}
$$

K.0.5 Revisiting Poisson's equation

The gravitational potential Poisson equation is $\nabla^2 U = 4\pi\mathcal{G}\rho$. This can and should be verified for the derived equations for prisms. Inside the prism, the density has an assigned constant value. Outside of the prism, the density is zero, so the result is $\nabla^2 U = 0$. These cases will be treated separately.

Outside the prism

$\nabla^2 U = 0$ can be written $\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0$. which is the trace of $\boldsymbol{T}$. We first add the terms and then the boundary conditions are applied.

We will need the formula to add arctangents together:

$$\arctan a + \arctan b = \arctan \frac{a+b}{1-ab} \tag{K.32}$$

We start by adding the terms T_{xx} and T_{yy} together:

$$
\begin{aligned}
T_{xx} + T_{yy} &= (-1)^{i+j+k}\arctan\frac{yz}{xr} + (-1)^{i+j+k}\arctan\frac{xz}{yr} \\
&= (-1)^{i+j+k}\arctan\frac{\frac{yz}{xr}+\frac{xz}{yr}}{1-\frac{yz}{xr}\frac{xz}{yr}} \\
&= (-1)^{i+j+k}\arctan\frac{\frac{y^2z}{xyr}+\frac{x^2z}{xyr}}{1-\frac{xyz^2}{xyr^2}} \\
&= (-1)^{i+j+k}\arctan\frac{\frac{z(x^2+y^2)}{xyr}}{\frac{xy(r^2-z^2)}{xyr^2}} \\
&= (-1)^{i+j+k}\arctan\frac{xyzr^2(x^2+y^2)}{x^2y^2r(r^2-z^2)} \\
&= (-1)^{i+j+k}\arctan\frac{zr(x^2+y^2)}{xy(x^2+y^2+z^2-z^2)} \\
&= (-1)^{i+j+k}\arctan\frac{zr}{xy} \tag{K.33}
\end{aligned}
$$

By considering a right triangle with sides 1 and x, it easy to prove that:

$$\arctan x + \arctan\frac{1}{x} = \frac{\pi}{2} \tag{K.34}$$

This can be used to transform the arctan to one that is similar to T_{zz}.

$$T_{xx} + T_{yy} = (-1)^{i+j+k}\arctan\frac{zr}{xy} = (-1)^{i+j+k}\left(\frac{\pi}{2} - \arctan\frac{xy}{zr}\right) \tag{K.35}$$

The last step is to add the term T_{zz}:

$$\nabla^2 U = T_{xx} + T_{yy} + T_{zz} = (-1)^{i+j+k}\left(\frac{\pi}{2} - \arctan\frac{xy}{zr} + \arctan\frac{xy}{zr}\right) = (-1)^{i+j+k}\frac{\pi}{2} \tag{K.36}$$

The end result is a single value. When the boundary conditions are applied this single value will be subtracted from itself resulting in zero, so $\nabla^2 U(\vec{r}) = 0$.

Inside the prism

We can simply put the observation point at the centre of the prism. The coordinates of the prism are now such that $-x_1 = x_2$, $-y_1 = y_2$ and $-z_1 = z_2$. All eight terms for these conditions results give $\frac{\pi}{2}$, so the result is:

$$\nabla^2 U = \mathcal{G}\rho 8\frac{\pi}{2} = 4\pi\mathcal{G}\rho \tag{K.37}$$

K.0.6 Better numerical stability

Heck and Seitz (2007) [590] modify the standard formulae for the prism to get a better numerical stability in the logarithms. This is done by dividing the inside of the logs by an extra factor:

$$\ln\left(z_k + r_{ijk}\right) \to \ln\frac{z_k + r_{ijk}}{\sqrt{x_i^2 + y_j^2}}$$

This extra factor disappears when applying the boundary conditions in the z direction, so that the results remain identical:

$$
\begin{aligned}
\left|\ln\frac{z_k + r_{ijk}}{\sqrt{x_i^2 + y_j^2}}\right|_{z_1}^{z_2} &= \left|\ln\left(z_k + r_{ijk}\right) - \ln\sqrt{x_i^2 + y_j^2}\right|_{z_1}^{z_2}\\
&= \ln\left(z_2 + r_{ijk}\right) - \ln\sqrt{x_i^2 + y_j^2} - \ln\left(z_1 + r_{ijk}\right) + \ln\sqrt{x_i^2 + y_j^2}\\
&= \ln\left(z_2 + r_{ijk}\right) - \ln\left(z_1 + r_{ijk}\right) \tag{K.38}
\end{aligned}
$$

Appendix L

A quick guide to Paraview and gnuplot

L.0.1   Paraview

Installation procedure

If you have Ubuntu, type the following in a terminal and follow the instructions.

```
sudo apt-get install paraview
```

Upon completion, type 'paraview' followed by Enter in the terminal and your screen should look similar to Screen Capture 1.

   If you run Windows or MacOS[1], go to `www.paraview.org`. Click on 'download'. The website automatically detects your OS[2]. Download the latest version(.exe for Windows, .dmg for Apple), and install it on your computer. Find the icon on your computer, double click on it and your screen should look similar to Screen Capture 1.
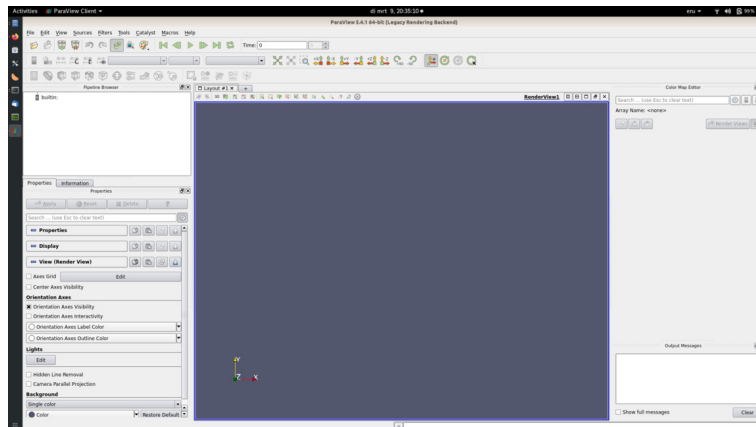
Screen Capture 1.

Opening a file

Press Ctrl+O on your keyboard or click File>Open> and the following window should appear after you press the green button Apply:

---

[1]You could also use Home Brew `https://formulae.brew.sh/cask/paraview`

[2]`https://en.wikipedia.org/wiki/Operating_system`

Left click of the mouse allows you to move the domain in the plotting area, Right click allows you to zoom in and out.

Select the .vtu file you wish to import in the session or a whole list of them. Your Paraview session should then look like this:

1: click on this icon to change the background colour to white;
2: click this icon for plotting vector field arrows and follow instructions below;
3: click this icon for isocontours and follow the intructions below.

4: click on this menu to select the field you wish to plot;
5: click on this menu and select Surface With Edges to see the mesh;
6: click on this icon to remove the red-yellow-green axis in the plotting area.

Colours and log scale

I have loaded the .vtu file, chosen a white background, zoomed in, selected the temperature field, so that my screen looks now like this (I have also moved the colour bar):

7: change the range of the variable;
8: change the colour scale;
9: change the number of colours inside the scale;

10: switch on logarithmic scale.

When it comes to choosing colours, please see: Crameri, Shephard, and Heron [305] and Zelst, Crameri, Pusok, Glerum, Dannberg, and Thieulot [1504].
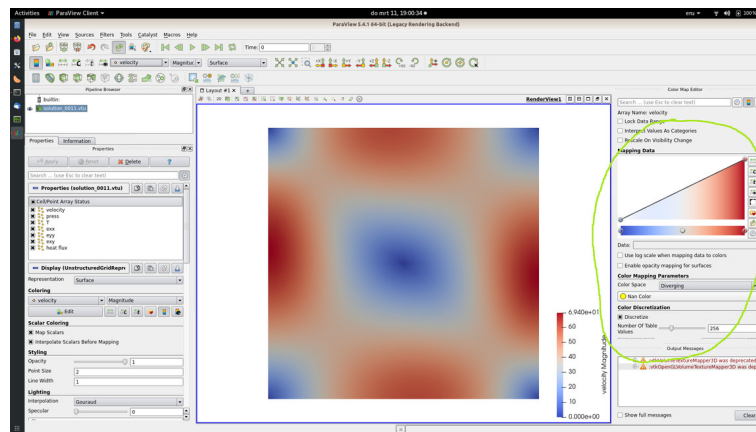
If you find that the circled area is missing on your screen, go to View and click on Color Map Editor.

> add how to add color scales

Isocontours

Having clicked on the icon numbered 3 in the panels above, your screen should look like this:
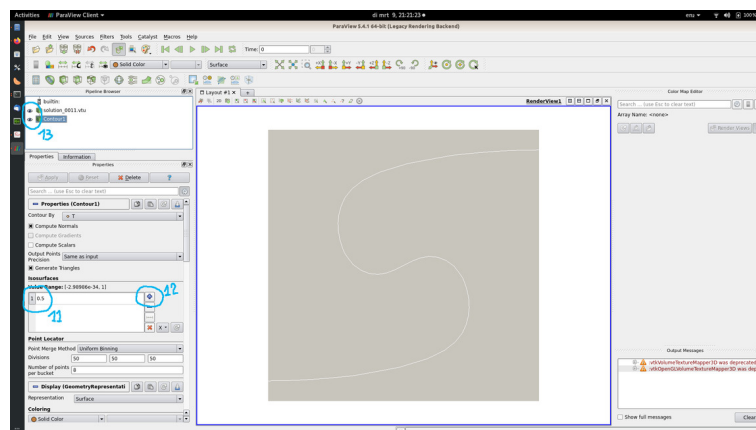
11: value of the isocontour;
12: add or remove an isocontour;

13: toggle the background grey square and the isocontours on/off by clicking on their respective eye.
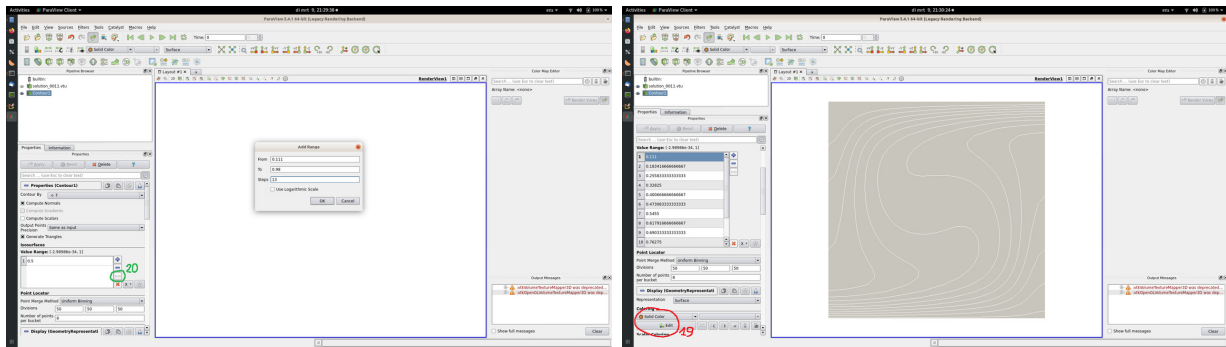
Left: If you want automatically generated isocontours, remove the existing one and click on 20. A small window opens: fill the min/max/number values and click OK. Right: Having obtained these isocontours you can change the colour of the lines by clicking on 19.

Vector field arrows
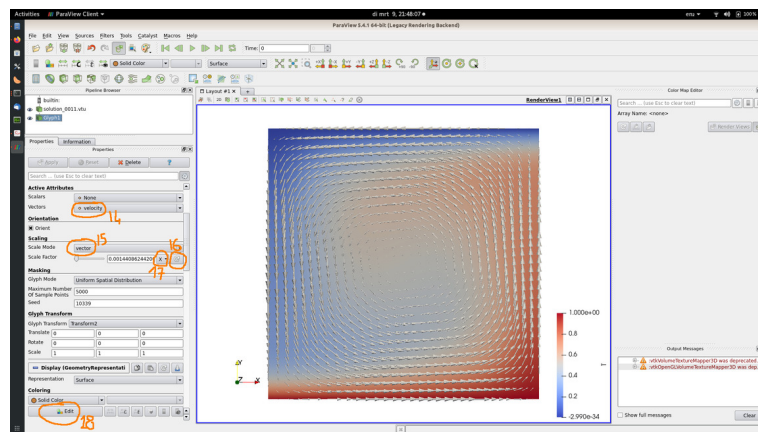
In order to obtain such arrows, make sure that you go through points 14 and 15.

Then click on the icon 16. In order to change the scale of the arrows change the value in 17.

Exporting to png

File>Save Screenshot. Click OK on the first panel. Enter the name of the file you have chosen and click OK.

Exporting line data

Filters > Data Analysis > Plot Over Line.
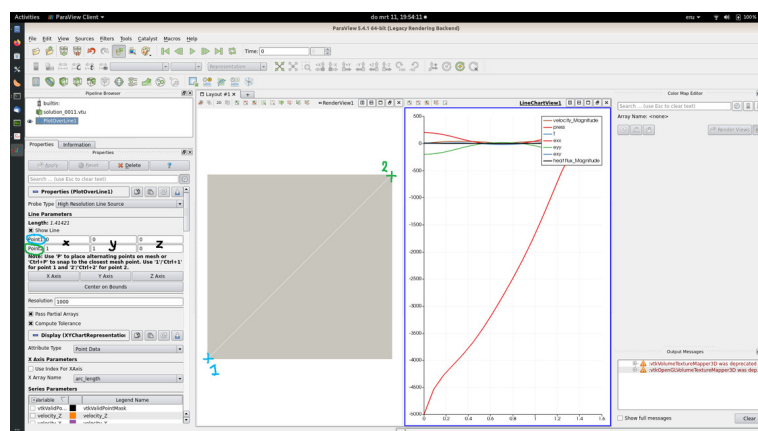
You can change the coordinates of the beginning and the end of the line.

Getting rid of 'Apply'

It can be annoying to have to press Apply all the time so if you wish to bypass it, got to Edit > Settings, and tick the 'Auto Apply' box in the window that appears.

Multiple vtu files at once
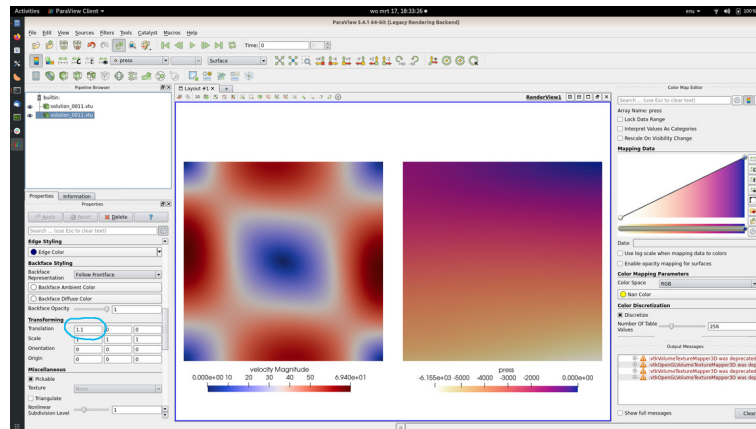
You can load multiple vtu files or the same one multiple times and move each where you want it.

Warp by scalar
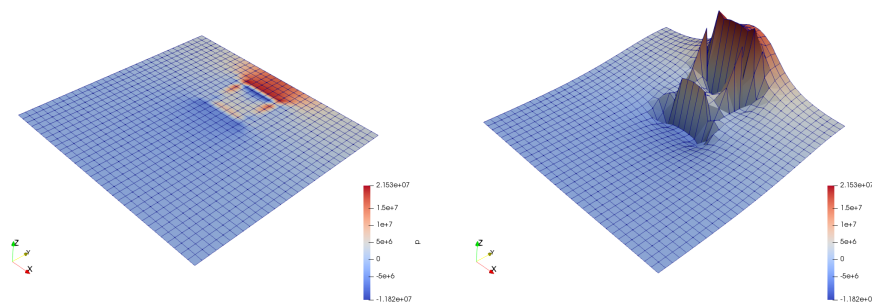
Filters > Alphabetical > Warp By Scalar

L.0.2 gnuplot

gnuplot is a famous and widely used command-line program that can generate two- and three-dimensional plots of functions, data, and data fits. It dates back to 1986 and runs on all operating systems (Linux, Unix, Microsoft Windows, macOS).

http://www.gnuplot.info/
http://www.gnuplotting.org/
http://lowrank.net/gnuplot/index-e.html
The gray boxes indicate that its content takes place in the terminal.

Installing gnuplot

If you are using Ubuntu, you can install gnuplot as follows:

```
> sudo apt-get install gnuplot
```

Interactive use

In the following pages I explain how to use the gnuplot program from the terminal. Having done so, in the terminal simply type

```
> gnuplot
```

The following should then appear:

```
G N U P L O T
Version 5.2 patchlevel 2     last modified 2017-11-01

Copyright (C) 1986-1993, 1998, 2004, 2007-2017
Thomas Williams, Colin Kelley and many others

gnuplot home:     http://www.gnuplot.info
faq, bugs, etc:   type "help FAQ"
immediate help:   type "help"  (plot window: hit 'h')

Terminal type is now 'wxt'

gnuplot>
```

The prompt means that gnuplot is expecting instructions. We start by making sure that the terminal type is such that a window appears in this interactive mode. We test this by plotting a simple function, $f(x) = x$:

```
gnuplot> plot x
```

You should then obtain something similar:
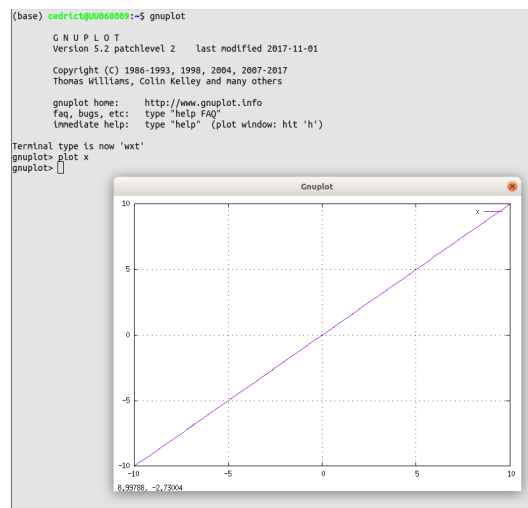
You can specify the x range, change the function to $x^2 + \sqrt{x}$ and label the axes as follows:

```
gnuplot> set xlabel 'time'
gnuplot> set ylabel 'cost'
gnuplot> plot[-5:7] x**2+sqrt(x)
```

1354
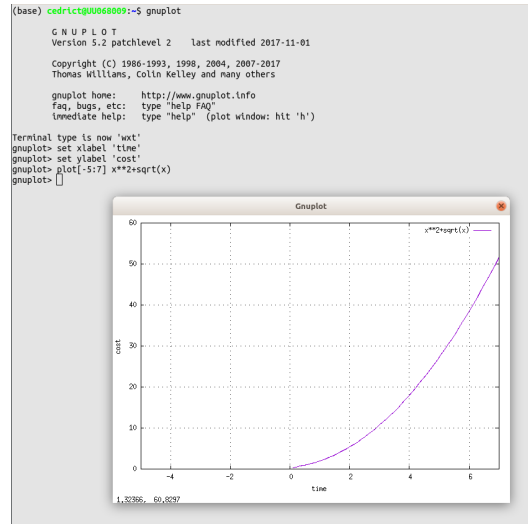
We can also plot functions of both x and y as follows:

```
gnuplot> splot x*y, x**2+y
```

and we get
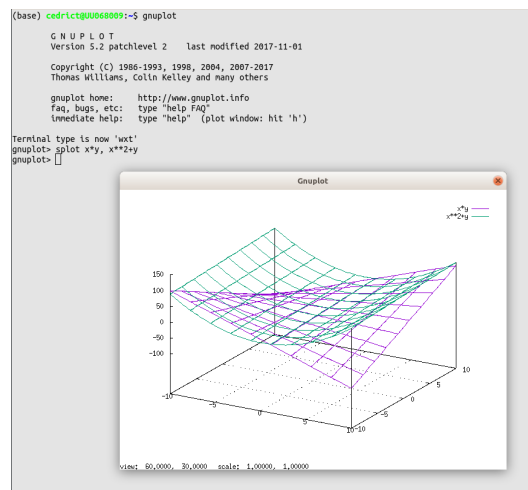
Another nice feature in the interactive is the fact that you can use the left button of the mouse to rotate the plot!

Finally, let us assume that there is a file *results.dat* in the folder and that it contains results from experimental measurements or numerical values organised in columns as follows:

```
1e17 8 0 256000 384000 4.91094e-12 -0.00533647 -714769 0
1e17 32 0 256000 384000 1.96438e-11 -0.0213459 -2.85908e+06 0
1e17 128 0 256000 384000 7.8575e-11 -0.0853835 -1.14363e+07 0
2e17 8 0 256000 384000 3.43871e-12 -0.00533555 -714753 0
2e17 32 0 256000 384000 1.37548e-11 -0.0213422 -2.85901e+06 0
2e17 128 0 256000 384000 5.50193e-11 -0.0853688 -1.1436e+07 0
4e17 8 0 256000 384000 4.13458e-12 -0.00533372 -714720 0
...
67108864e17 128 0 256000 384000 -5.28841e-12 -0.0212269 -1.27701e+07 0
134217728e17 8 0 256000 384000 2.93622e-13 -0.00132619 -798163 0
134217728e17 32 0 256000 384000 1.17449e-12 -0.00530475 -3.19265e+06 0
134217728e17 128 0 256000 384000 4.69795e-12 -0.021219 -1.27706e+07 0
```

```
268435456e17 8 0 256000 384000 4.03077e-13 -0.00132594 -798181 0
268435456e17 32 0 256000 384000 1.61231e-12 -0.00530376 -3.19272e+06 0
268435456e17 128 0 256000 384000 6.44923e-12 -0.0212151 -1.27709e+07 0
```

In this case we may with to plot the 6th column as a function of the 1st one:

```
plot 'results.dat' using 1:6 with linespoint linewidth 2 title 'velocity'
```
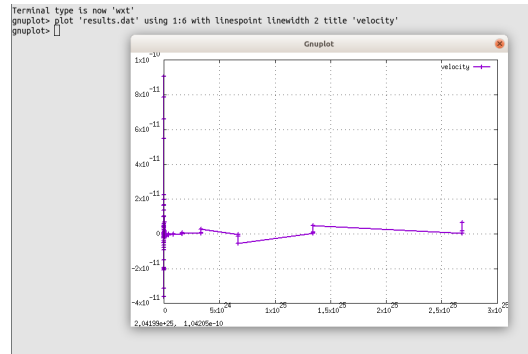

Since typing these instructions time and time again is a bit tedious gnuplot allows the user to use short versions of these commands:

```
gnuplot> plot 'results.dat' u 1:6 w lp lw 2 t 'velocity'
```

We see that the range of x values spans many order of magnitudes so we wish to use a logarithmic scale on the x-axis.

```
gnuplot> set log x
```

Also, I can combine data with analytical function:

```
gnuplot> set log x
gnuplot> plot 'results.dat' u 1:6 w lp lw 2 t 'velocity', 1e7/x lw 3 , 6e-11
```
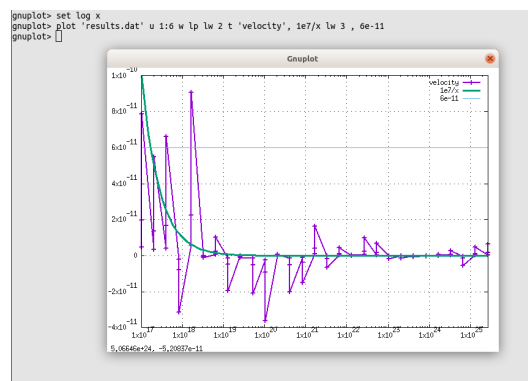

Finally, we may wish to export the plot to a file, say a pdf file. We must then re-assign the terminal, give a name to the file and re-plot:

```
gnuplot> set term pdf
gnuplot> set output 'results.pdf'
gnuplot> plot 'results.dat' u 1:6 w lp lw 2 t 'velocity', 1e7/x lw 3 , 6e-11
```

You can exit the session by typing

```
gnuplot> exit
```

You should find *results.pdf* in your folder next to *results.dat*.

Scripting gnuplot

Although the interactive approach is very useful its workflow is not practical if one wishes (for instance) to produce the same plot with different data, or to communicate a figure to another scientist.

We will therefore now turn to scripting. The idea is simple: write all the gnuplot commands in a text file, say *script.gnuplot* and pass this script as argument to gnuplot:

```
> gnuplot script.gnuplot
```

This file contains the following:

```
set term pdf font "Times,12pt"
set output 'results.pdf'
set grid
set xlabel 'x'
set ylabel 'cost'
set log x
set title 'my title above the plot'
plot 'results.dat' u 1:6 w lp lw 2 t 'velocity', 1e7/x lw 3 t 'fit' , 6e-11 t 'threshold'
```
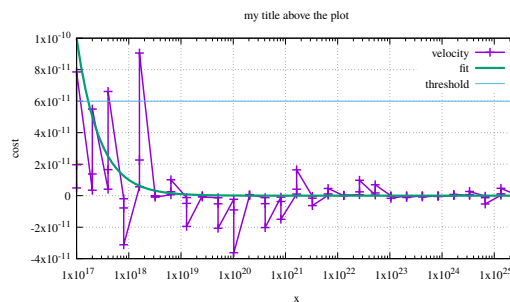

Note that I have added a title to the plot as well.

Greek letters

In order to display Greek letters the `/Symbol` command:

```
set xlabel '{/Symbol d}{/Symbol r}'
```

Alphabet	Symbol	Alphabet	Symbol	Alphabet	Symbol	Alphabet	Symbol
A	Alpha	N	Nu	a	alpha (α)	n	nu ν
B	Beta	O	Omicron	b	beta (β)	o	omicron
C	Chi	P	Pi	c	chi (χ)	p	pi π
D	Delta	Q	Theta	d	delta (δ)	q	theta θ
E	Epsilon	R	Rho	e	epsilon (ϵ)	r	rho ρ
F	Phi	S	Sigma	f	phi (ϕ)	s	sigma σ
G	Gamma	T	Tau	g	gamma (γ)	t	tau τ
H	Eta	U	Upsilon	h	eta (η)	u	upsilon υ
I	iota	W	Omega	i	iota (ι)	w	omega ω
K	Kappa	X	Xi	k	kappa (κ)	x	xi ξ
L	Lambda	Y	Psi	l	lambda (λ)	y	psi ψ
M	Mu	Z	Zeta	m	mu (μ)	z	zeta ζ

piecewise function

You can define piecewise functions as follows:

```
f(x) = x<a  ? 1 : 1/0
g(x) = x>=a ? 1 : 1/0
```

and then use these functions like any function, e.g.:

```
plot[-10:12] f(x),g(x)
```

linetype and dashtype

There are essentially three ways of plotting data:

```
plot 'results.dat' u 1:6 w l
plot 'results.dat' u 1:6 w p
plot 'results.dat' u 1:6 w lp
```

corresponding to (from left to right):
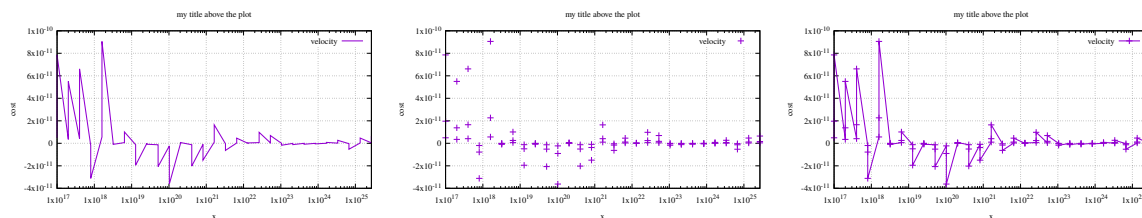
The following script

```
set output 'linetypes.pdf'
plot[][]\
x+0  w l lt 0 t 'linetype 1',\
x+1  w l lt 1 t 'linetype 2',\
x+2  w l lt 2 t 'linetype 3',\
x+3  w l lt 3 t 'linetype 4',\
x+4  w l lt 4 t 'linetype 5',\
x+5  w l lt 5 t 'linetype 6',\
x+6  w l lt 6 t 'linetype 7',\
x+7  w l lt 7 t 'linetype 8',\
x+8  w l lt 8 t 'linetype 9',\
x+9  w l lt 9 t 'linetype 10',\
x+10 w l lt 10 t 'linetype 11',\
x+11 w l lt 11 t 'linetype 12'
```

generates the following plot:

1358
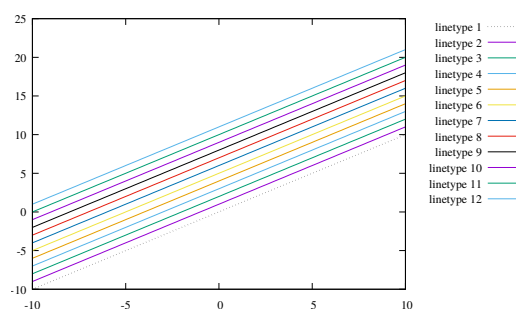
We see that the colours repeat from linetype 10. Fortunately we can also combine linetypes with dashtypes. The following script

```
set output 'dashtypes.pdf'
plot[][]\
x+1  w l lt 1  dt 1 t 'linetype 2',\
x+2  w l lt 2  dt 2 t 'linetype 3',\
x+3  w l lt 3  dt 3 t 'linetype 4',\
x+4  w l lt 4  dt 4 t 'linetype 5',\
x+5  w l lt 5  dt 5 t 'linetype 6',\
x+6  w l lt 6  dt 6 t 'linetype 7',\
x+7  w l lt 7  dt 7 t 'linetype 8',\
x+8  w l lt 8  dt 8 t 'linetype 9',\
x+9  w l lt 9  dt 9 t 'linetype 10',\
x+10 w l lt 10 dt 10 t 'linetype 11',\
x+11 w l lt 11 dt 11 t 'linetype 12'
```

generates the following plot
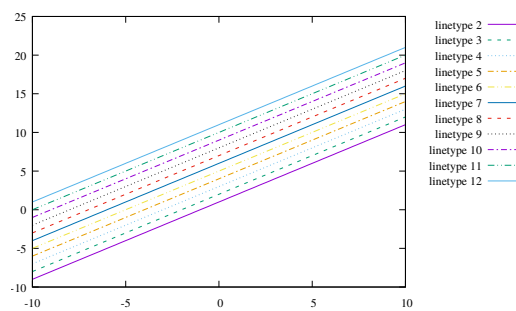
and we see that there are only 5 different dash types.

Finally, we turn to point types. The following script

```
set output 'pointtypes.pdf'
plot[][]\
x+0  w p pt 0  ps .5 t 'linetype 1',\
x+1  w p pt 1  ps .5 t 'linetype 2',\
x+2  w p pt 2  ps .5 t 'linetype 3',\
x+3  w p pt 3  ps .5 t 'linetype 4',\
x+4  w p pt 4  ps .5 t 'linetype 5',\
x+5  w p pt 5  ps .5 t 'linetype 6',\
x+6  w p pt 6  ps .5 t 'linetype 7',\
x+7  w p pt 7  ps .5 t 'linetype 8',\
x+8  w p pt 8  ps .5 t 'linetype 9',\
x+10 w p pt 10 ps .5 t 'pointtype 10',\
x+11 w p pt 11 ps .5 t 'pointtype 11',\
x+12 w p pt 12 ps .5 t 'pointtype 12',\
x+13 w p pt 12 ps .5 t 'pointtype 13',\
x+14 w p pt 12 ps .5 t 'pointtype 14'
```

generates the following plot
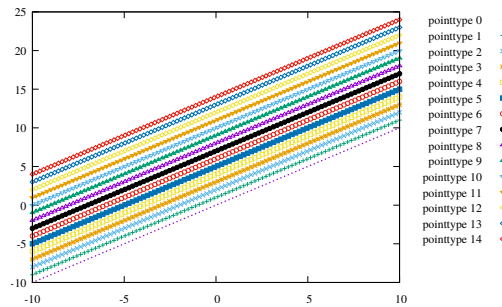
Note that I have used the `ps` command ('point size') to make the points 50% smaller than normal.

Moving the 'key'

The default is inside top right, but it can be changed, e.g.:

```
set key outside
set key bottom left
```

corresponding to (from left to right):
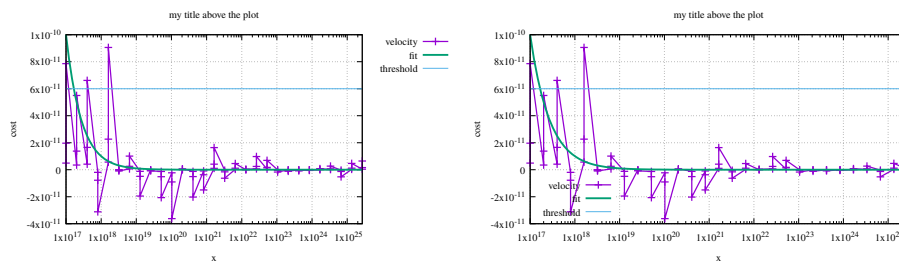
Plotting arrows

Let us now turn to the *velocity.dat* file which consists of four columns: x, y, ν_x and ν_y.

```
set output 'velocity_1.pdf'
set xlabel 'x'
set ylabel 'y'
set xtics 0.125
set ytics 0.333333333333
set grid
set size square
plot[0:1][0:1]\
'velocity.dat' u 1:2:3:4 w vectors lt -1 notitle
```

Note that I have required the plot to be square, that the tics on the x-axis should be spaced 0.125 while the tics on the y-axis should be spaced 0.333. We obtain the left plot a):
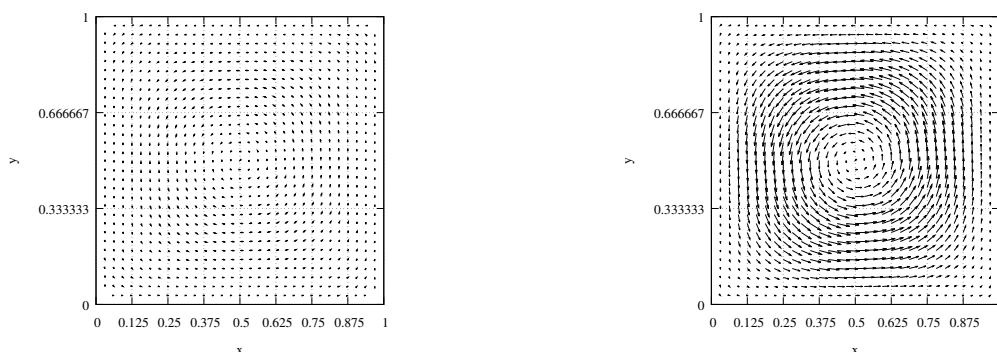
The arrows are too small, so we scale each vector component by a factor 4. All we need to do is as follows:

```
plot[0:1][0:1]\
'velocity.dat' u 1:2:($3*4):($4*4) w vectors lt -1 notitle
```

Note the dollar sign which means that gnuplot takes the value in column 3 or 4 and multiplies it by 4. The resulting figure is shown in b).

Powers of 10

```
set format y "10^{%L}"
```

Least square fit

Assuming we have a file containing data, e.g. `data.ascii`, that we want to fit by means of a linear relationship over the range $[-1, +1]$:

```
f(x)=a*x+b
fit [-1:1] f(x) 'data.ascii' u 1:2 via a,b
```

This should return a few lines in the console indicating whether convergence was reached and then also the a and b values. In order to plot the line, simply do:

```
plot[] 'data.ascii' u 1:2, f(x)
```

coloring areas

```
set style rect fc lt -1 fs solid 0.1 noborder
set obj rect from 0, graph 0 to 15, graph 1
```

vertical line

To draw a vertical line from the bottom to the top of the graph at x=3, use:

```
set arrow from 3, graph 0 to 3, graph 1 nohead
```

Show list of all available colors

In an interactive gnuplot session type:

```
> show colors
```

Piecewise functions

```
f(x) = x<a ? 1: 1/0
g(x) = x>=a ? 1: 1/0
```

Note that the 1/0 results in an infinity which will not plot.

Appendix M

A few LATEXfeatures

newcommand

This features allows to define new commands which can then be used throughout the document. In *manual.tex* you will find

```
\newcommand{\K}{{\mathbb{K}}}
\newcommand{\Ranb}{{\mathsf{Ra}}}
\newcommand{\nineteeneightysix}{{\color{violet}\bf 1986}}
```

which correspond to $\mathbb{K}$, Ra and **1986**.

fullpage package

How to extend margins for the whole document (such as this one):

```
\usepackage[cm]{fullpage}
```

siunitx package

```
\usepackage{siunitx}
\DeclareSIUnit\year{yr}
```

tikz

```
\usetikzlibrary{arrows, arrows.meta}
```

include verbatim material inside a line

`\verb"git pull upstream master"` and then eat an ice cream.

results in `git pull upstream master` and then eat an ice cream.

bibliography

opla
```
\footfullcite
```

how to refer to a latex document from another document?

This is the first latex file names `manual.tex`. It contains sections and equations, all labelled.

```
\documentclass[a4paper]{book}
\begin{document}
\chapter{opla1} \label{ch1}
\section{meuh} \label{ss1}
\subsection{popo}
opla \ref{ch1} opla \ref{ss1}
\begin{equation}
\alpha= \beta \label{eq:one}
\end{equation}
\end{document}
```

This second file will load the `manual.aux` with the 'xr' package and then all labels of that file are now available in this file:

```
\documentclass[a4paper]{article}
\usepackage{amsmath}

\usepackage{xr}
\externaldocument[MMM-]{manual}

\begin{document}

opla

the introduction to volume1 (\ref{MMM-ss1})
see \eqref{MMM-eq:one}

\end{document}
```

Appendix N

Linux how to

When encountering the terminal for the first time one soon realises that some basic commands are needed to navigate the folders, edit, copy or delete files, etc ...

Please also have a look at this website: `https://ryanstutorials.net/linuxtutorial/`

For Windows users, you must abandon your preconceived (and totally arbitrary) ideas about the 'C:' drive. Before hard drives even existed, the computer would have one or two floppy drives and would reserve the A: and B: drive letters for them. Nowadays these devices are usually not installed in the computer anymore, but the labelling remains[1].

The root of the file system is simply /. Your 'home' is most likely in `/home/your-family-name/`. Inside this you will find `Documents`, `Downloads`, etc ...

The basic set of commands is in fact not very large (Please also watch `https://youtu.be/6bMYzzrycV0`):

- **man**: Documentation in Linux is mostly available in the form of *man pages*. They are usually written in the style of a reference manual and can be daunting at first. They are usually: the name, a compact formulation of the syntax (that can be scary for more complex programs), a description about what the software actually is and does, examples (if you are lucky) and explanations of all the options mentioned above. Typical use:

  ```
  > man name-of-command-I-want-to-learn-about
  ```

- **cd**: it stands for 'change directory' (i.e. change folder). Typical use:

  ```
  > cd results
  ```

  where `results` should be an existing folder. You can check this with:

- **ls**: it stands for 'list'. This commands lists all files and folders Typical use:

  ```
  > ls
  ```

  or

  ```
  > ls -la
  ```

  if you wish to have one item per line, or

---

[1] `https://en.wikipedia.org/wiki/Drive_letter_assignment`

```
> ls -l
```

if you also want to see all files/folders starting with '.' My preference goes to

```
> ls -lhG
```

Use the 'man' command to know what this does!

- **mv**: it stands for 'move' but it has in fact a hidden functionality: rename. If you type

```
> mv garfield.txt odie.txt
```

then the file *garfield.txt* has been renamed *odie.txt* [2].

You can move a file in a different folder as follows:

```
> mv garfield.txt ../myfolder/
```

This moves the file to a folder that is one level up. This will work only if the folder exists. If you need to make a new folder, then use:

- **mkdir**: it stands for 'make directory'. Typical usage

```
> mkdir res_123
```

creates a folder named *res_123*.

- **rm**: it stands for 'remove'. Before we go any further: this command is dangerous. Unlike its counterpart based on selecting a file with a mouse and deleting it, rm does not send the file to the Trash folder (or Windows Recycle Bin). It simply deletes it forever. No turning back! Typical usage

```
> rm myoldfile.txt
```

deletes the file. Note that by default, it does not remove directories. In order to remove a folder, one needs to type

```
> rm -r myoldfolder
```

Unless you are an experienced user, never use `rm` in conjunction with * and/or in a recursive way. If things go wrong, you will delete entire portions of your hard drive at best, or will destroy your operating system at worse.

- **pwd**: it stands for 'print working directory'. If you are unsure of where the current prompt of the terminal is, simply

```
> pwd
```

and it will return the full path, from the root to where the prompt is.

- **du**: it stands for 'disk usage'. In this case always tag the -h option to it:

---

[2]`https://en.wikipedia.org/wiki/Odie` Obviously one cannot transform Garfield into Odie.

```
> du -h
```

It will list the size of all folders in the folder the prompt is in. If you wish to know the size of an object simply do

```
> du -h file-or-folder
```

The `ls -lh` command would have told you as much, but for all files inside the folder.

- **grep**: searches for patterns in each file. Typical usage

```
> grep linear *.tex
```

This searches all occurrences of the word 'linear' in all .tex files. If you wish to look for this word in all files inside subfolders:

```
> grep -r linear .
```

- **more**: allows to visualise the content of an ascii[3] file inside the terminal without using a text editor. In other words, you can look into the file but not change its content.

```
> more interesting-file
```

- **top/htop**: allows to visualise which process is running on the computer and how much CPU and memory it takes.

- **wget**: it is a computer program that retrieves content from web servers. It is part of the GNU Project. Its name derives from "World Wide Web" and "get." It supports downloading via HTTP, HTTPS, and FTP.

```
> wget address.of.file.online.on.server
```

See example of use in Section 18.12.5.

- **ssh**: This command is necessary to connect to a remote computer from the terminal. By default most Linux computers run a client and server ssh program so that one can connect to any other computer if its address is known (as well as the username and password). Here is how I connect to my shrek desktop computer:

```
> ssh shrek.geo.uu.nl -Y -l thieulot
```

The -Y option ensures that I have X11 support. If successful, the prompt of the terminal points to the default folder on the remote machine and I can control the remote computer via the command line. Type exit to break the connection.

This is used to log in on remote servers like clusters where large calculations take place.

- **sftp**: allows to get or put files on a remote computer via a secrured connection. By default, SFTP uses the SSH protocol to authenticate and establish a secure connection. Because of this, the same authentication methods are available that are present in SSH.

---

[3]https://en.wikipedia.org/wiki/ASCII

```
> sftp garfield@remote.univ.nl
```

This will prompt for the password of the remote machine associated to user garfield. If succesful the prompt in the terminal then points to the default folder on the remote computer. Once connected simply type exit to break the connection. All shell commands are available: pwd, ls, cd, etc ... you can list the contents of the current directory on the local machine using `lls`. If we want to download files from our remote host, you can do so using the get command:

```
> get remotefile
```

If you wish to download an entire folder, simply use

```
> get -r remotefolder
```

Conversely, you can transfer files from your local machine onto the remote one:

```
> put remotefile
```

- **scp**: to do

In what follows I list a few 'tricks' which I find useful or just can never remember:

- How to convert files in batch

```
> convert '*.png' converted_%04d.jpg
```

- How to Remove unwanted empty lines in a file with vi(m)
  Use either of the following commands to delete all empty lines:

```
:g/^$/d
:v/./d
```

  If you want to delete all lines that are empty or that contain only whitespace characters (spaces, tabs), use either of:

```
:g/^\s*$/d
:v/\S/d
```

- How to find LAPACK

```
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
```

- How to apt-get MUMPS

```
> sudo apt-get install libmumps-seq-dev
```

- How to remove a big file wrongly committed

1367

```
> git filter-branch --tree-filter 'rm -rf path/to/your/file' HEAD
> git push
```

- How to check your disk usage (ubuntu)

  Disk Usage Analyzer (aka Baobab) is a graphical, menu-driven viewer that you can use to view and monitor your disk usage and folder structure. It is part of every GNOME desktop.

  ```
  > baobab
  ```
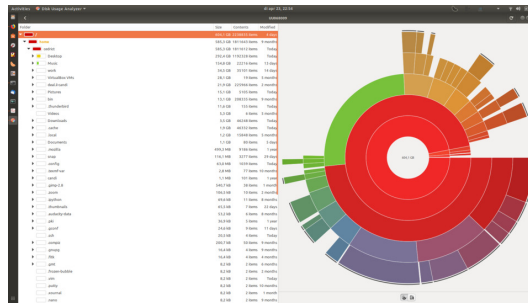

- How to convert images into another format in the command line

  ```
  > convert file.png file.jpg
  ```

  You can also resize on the fly:

  ```
  > convert -resize 70% file.png file.jpg
  ```

- How to compress files: tar is used to pack files into an archive without zipping them. This enables easy attaching to email, or to sending across the internet without the computer having to start a new transfer for every file. Also it ensures everything belong to a certain package stays together. Optionally it can be used to zip files. Let us have an example of three files; *a.txt, b.py* and *c.dat* These can be packed together with

  ```
  > tar -cf allFiles.tar *
  ```

  c stands for *compress* and f for *file*, as can be seen in the manpage of tar. It can similarly be extracted using the command:

  ```
  > tar -xf allFiles.tar *
  ```

  to recover three original files.

  Additionally, it is possible to zip them as well, by adding a z to the options. The default extension then becomes *tgz*:

  ```
  > tar -czf allFiles.tgz *
  > tar -xzf allFiles.tgz *
  ```

1368

Appendix O

on using Fortran

Cool video: Fortran in 100s `https://youtu.be/NMWzgy8FsKs?si=KekMuUVEm6kt1zIe`

O.0.1 Full matrix multiplications in fortran

In fortran there is the intrinsic function *matmul*. However, it turns out that it is not always the fastest option to carry out (full) matrix multiplications.

This code is designed to test this:

```fortran
program test
implicit none
! The order of the square matrices is 2000.
integer(kind=4)::n=1000
! Calculate the matrix multiplications:
! i) c:=a*b in a triple do-loop.
! ii) d:=a*b by matmul(a,b).
! iii) e:=a*b by dgemm in INTEL MKL.
real(kind=8),allocatable::a(:,:),b(:,:),c(:,:),d(:,:),e(:,:)
real(kind=8)::alpha,beta
integer(kind=4)::i,j,k,lda,ldb,lde
real(kind=8)::start,finish

allocate(a(n,n),b(n,n),c(n,n),d(n,n),e(n,n))
alpha=1.0;beta=1.0
lda=n;ldb=n;lde=n

! Generate the matrices, a and b, randomly.
call cpu_time(start)
call random_seed()
do j=1, n
do i=1, n
call random_number(a(i,j))
call random_number(b(i,j))
enddo
enddo
call cpu_time(finish)
write(unit=6,fmt=100) "The generation of two matrices takes ",finish-start," seconds.
    "

! i) c:=a*b in a triple do-loop.
call cpu_time(start)
c=0.0D0
do j=1, n
do i=1, n
do k=1, n
```

```fortran
c(i,j)=c(i,j)+a(i,k)*b(k,j)
enddo
enddo
enddo
call cpu_time(finish)
write(unit=6,fmt=100) "A triple do-loop takes ",finish-start,"seconds."

! ii) d:=a*b by matmul(a,b).
call cpu_time(start)
d=0.0D0
d=matmul(a,b)
call cpu_time(finish)
write(unit=6,fmt=100) "A matmul(a,b) function takes ",finish-start,"seconds."

! iii) e:=a*b by dgemm in INTEL MKL.
call cpu_time(start)
e=0.0D0
call dgemm("N","N",n,n,n,alpha,a,lda,b,ldb,beta,e,lde)
call cpu_time(finish)
write(unit=6,fmt=100) "A DGEMM subroutine takes ",finish-start,"seconds."

deallocate(a,b,c,d,e)

stop
100 format(A,F8.3,A)
end program test
```

It is compiled as follows:

```
> gfortran -O3 prog.f90 -lblas
```

For 100×100 matrices:

```
The generation of two matrices takes    0.004 seconds.
A triple do-loop takes    0.009 seconds.
A matmul(a,b) function takes    0.001 seconds.
A DGEMM subroutine takes    0.000 seconds.
```

For 1000×1000 matrices:

```
The generation of two matrices takes    0.123 seconds.
A triple do-loop takes    1.527 seconds.
A matmul(a,b) function takes    0.080 seconds.
A DGEMM subroutine takes    0.054 seconds.
```

For 1000×2000 matrices:

```
The generation of two matrices takes    0.392 seconds.
A triple do-loop takes   33.785 seconds.
A matmul(a,b) function takes    0.725 seconds.
A DGEMM subroutine takes    0.455 seconds.
```

O.0.2   A simple example of an Interface

```fortran
program kwadraat

implicit none
```

```fortran
integer, parameter :: IntegerRoot = 6
real,     parameter :: RealRoot = 4.5

Interface Square
    function RealSquare(root)
        real :: root
        real :: RealSquare
    end function

    function IntegerSquare(root)
        integer :: root
        integer :: IntegerSquare
    end function
end interface


write(*,*) "Integer_square:_", Square(IntegerRoot)
write(*,*) "Real_square:____", Square(realRoot)

end program

function IntegerSquare(root)
    implicit none
    integer :: root, IntegerSquare
    IntegerSquare = root**2
end function

function RealSquare(root)
    implicit none
    real :: root, RealSquare
    RealSquare = root*root
end function
```

Appendix P

mineral parameters

P.0.1 Olivine

ref.	Wet A	Q (kJ/mol)	V	n	Dry A	Q (kJ/mol)
[534]	$3.91 \cdot 10^3 MPa^{-n}/s$ $= 3.91 \cdot 10^{-15} Pa^{-n}/s$	430	0	3	$2.42 \times 10^5 MPa^{-n}/s$ $= 2.42 \times 10^{-16} Pa^{-n}/s$	540
[314]	$3.91 \cdot 10^{-15} Pa^{-n}/s$	430	0	3		
[646]					$2.4 \times 10^{-16} Pa^{-n}/s$	540
[647]			0		$2.42 \times 10^{-15} Pa^{-n}/s$	540
[714]	$3.906 \cdot 10^{-15} Pa^{-n}/s$	430	10-20	3	$2.4169 \cdot 10^{-16} Pa^{-n}/s$	540
[666]			0		$1.43 \times 10^{-15} Pa^{-n}/s$	65
[511]	$4.89 \cdot 10^{-15} Pa^{-n}/s$	515		3.5		
[1088]	$4.89 \cdot 10^{-15} Pa^{-n}/s$	515		3.5	$4.85 \times 10^{-17} Pa^{-n}/s$	535
[610]	$4.89 \cdot 10^{-15} Pa^{-n}/s$	515		3.5	$4.85 \times 10^{-17} Pa^{-n}/s$	535
[713]					$10^{6.1} MPa^{-n}/s$ $= 1.26 \cdot 10^{-12} Pa^{-n}/s$	510 ± 30
[1108]	$2 \times 10^3 MPa^{-n}/s$ $= 2 \cdot 10^{-21} Pa^{-n}/s$	471 ± 31		4 ± 0.1	$2.5 \times 10^4 MPa^{-n}/s$ $= 2.5 \cdot 10^{-17} Pa^{-n}/s$	532 ± 52
[1331]	$5.33 \cdot 10^{-19}$	480	11	3.5		
[1331]	$1.5 \cdot 10^{-18}$	335	4	1		
[726]					$10^{5.04} MPa^{-n}/s$ $= 1.1 \cdot 10^{-16} Pa^{-n}/s$	530
[828]						470
[611]	$3.58 \cdot 10^{-16} Pa^{-n}/s$	480 ± 40	11	3.5 ± 0.3	$1.1 \cdot 10^{-16} Pa^{-n}/s$	530 ± 4
[611]	$8 \cdot 10^{-9} Pa^{-n}/s$	335 ± 75	4	1	$1.2 \cdot 10^{-8} Pa^{-n}/s$	375 ± 50
ELEFANT						
wetolivine1	$3.9063 \cdot 10^{-15} Pa^{-n}/s$	430	15	3		
dryolivine1					$2.4169 \cdot 10^{-16} Pa^{-n}/s$	540
wetolivine2	$4.89 \cdot 10^{-15} Pa^{-n}/s$	515		3.5		
dryolivine2					$4.85 \times 10^{-17} Pa^{-n}/s$	535

P.0.2 Quartz

ref.	Wet A	Q (kJ/mol)	V ()	n	Dry A	Q (kJ/mol)	V ()	n
[534]	$3.2 \times 10^{-4} MPa^{-n}/s$ $= 5.072 \cdot 10^{-18} Pa^{-n}/s$	154		2.3				
[1331]	$8.57 \cdot 10^{-28} Pa^{-n}/s$ $\to 1.1 \cdot 10^{-28} Pa^{-n}/s$	223	0	4				
[666]	$8.574 \times 10^{-28} Pa^{-s}/s$ $\to 1.1 \cdot 10^{-28} Pa^{-n}/s$	26.8	0	4				
[646, 647, 314, 511]	$1.10 \times 10^{-28} Pa^{-s}/s$	223	0	4				
[66]	2.91×10^{-3}	151		1.8				
[494]	$1.8 \cdot 10^{-8\pm2} MPa^{-n}/s$ $= 1.8 \cdot 10^{-32\pm2} Pa^{-n}/s$	137 ± 34	0	4 ± 0.9				
	$1.1 \cdot 10^{-4\pm2} MPa^{-n}/s$ $= 1.1 \cdot 10^{-28\pm2} Pa^{-n}/s$	223 ± 56	0	4 ± 0.9				
ELEFANT `wetquartz1`	$1.1 \cdot 10^{-28\pm2} Pa^{-n}/s$	223 ± 56	0	4 ± 0.9				

note that Buiter in [1331] says that 8.57 value is alreay scaled. Same for [666]. There are indeed, because for $n = 4$ the multiplicative factor is approx. 7.794, and 8.574/7.794=1.1 as in [494].

P.0.3 Plagioclase

ref.	Wet A	Q (kJ/mol)	V ()	n	Dry A	Q (kJ/mol)	V ()	n	comm
[1107]					$3.2 \times 10^{-4} MPa^{-n}/s$ $= 3.2 \times 10^{-23.2} Pa^{-n}/s$ $= 2.02 \times 10^{-23} Pa^{-n}/s$	238 238 238		3.2 3.2 3.2	
ELEFANT `dryplagioclase1`					$2.02 \times 10^{-23} Pa^{-n}/s$	238		3.2	

P.0.4 Peridotite

ref.	Wet A	Q (kJ/mol)	V ()	n	Dry $A(Pa^{-n})$	Q (kJ/mol)	V ()	n	comment
[1107]	2.0×10^3	471		4	2.5×10^4	532		3.5	
ELEFANT `peridotite`	2.0×10^3	471		4	2.5×10^4	532		3.5	

P.0.5 Diabase

ref.	Wet A	Q (kJ/mol)	V ()	n	Dry $A(Pa^{-n})$	Q (J/mol)	V ()	n	comment
[314, 511]					5.04×10^{-28}	485		4.7	refers to [866]
[866]						485 ± 30		4.7 ± 0.6	
ELEFANT									

P.0.6 Gabbro

ref.	A	Q (kJ/mol)	V ()	n		comment
[1331]	$1.12 \cdot 10^{-10} Pa^{-n}/s$	497	0	3.4		refers to [1458]
ELEFANT						

Looking in [1458] , can't find the number !?!

P.0.7 Serpentine

ref.	A	Q (kJ/mol)	V ()	n		comment
[603]	$4.47 \cdot 10^{-38} Pa^{-n}/s$	8.9	$3.2 cm^3$	3.8		

Appendix Q

Invariants

Remember: $\mathcal{I}_{1,2,3}$ are moment invariants while $\mathcal{K}_{1,2,3}$ are principal invariants.

Second invariants

Remembering that the deviatoric stress tensor $\boldsymbol{\tau}$ is symmetric:

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\boldsymbol{\tau}:\boldsymbol{\tau} \\
&= \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2 + 2\tau_{xy}^2 + 2\tau_{xz}^2 + 2\tau_{yz}^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\sum_{ij}\tau_{ij}\tau_{ji} \\
&= \frac{1}{2}\sum_{ij}\tau_{ij}\tau_{ij} \qquad (\boldsymbol{\tau} \text{ is symm}) \\
&= \frac{1}{2}\boldsymbol{\tau}:\boldsymbol{\tau}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\text{tr}[\boldsymbol{\tau}\cdot\boldsymbol{\tau}] \\
&= \frac{1}{2}\text{tr}\left[\begin{pmatrix} \tau_{xx}^2 + \tau_{xy}\tau_{yx} + \tau_{xz}\tau_{zx} & \cdot & \cdot \\ \cdot & \tau_{yx}\tau_{xy} + \tau_{yy}^2 + \tau_{yz}\tau_{zy} & \cdot \\ \cdot & \cdot & \tau_{zx}\tau_{xz} + \tau_{zy}\tau_{yz} + \tau_{zz}^2 \end{pmatrix}\right] \\
&= \frac{1}{2}\text{tr}\left[\begin{pmatrix} \tau_{xx}^2 + \tau_{xy}^2 + \tau_{xz}^2 & \cdot & \cdot \\ \cdot & \tau_{xy}^2 + \tau_{yy}^2 + \tau_{yz}^2 & \cdot \\ \cdot & \cdot & \tau_{xz}^2 + \tau_{yz}^2 + \tau_{zz}^2 \end{pmatrix}\right] \qquad (\boldsymbol{\tau} \text{ is symm}) \\
&= \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2 + 2\tau_{xy}^2 + 2\tau_{xz}^2 + 2\tau_{yz}^2) \tag{Q.1}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{K}_2(\boldsymbol{\sigma}) \;&=\; \frac{1}{2}[\mathrm{tr}(\boldsymbol{\sigma})^2 - \mathrm{tr}(\boldsymbol{\sigma}^2)] \\[2mm]
&=\; \frac{1}{2}\left[(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})^2 - (\sigma_{xx}^2 + \sigma_{xy}\sigma_{yx} + \sigma_{xz}\sigma_{zx} + \sigma_{yy}^2 + \sigma_{xy}\sigma_{yx} + \sigma_{yz}\sigma_{zy} + \sigma_{zz}^2 + \sigma_{xz}\sigma_{zx} + \sigma_{yz}\sigma_{zy})\right] \\[2mm]
&=\; \frac{1}{2}\left[(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})^2 - (\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xy}\sigma_{yx} + 2\sigma_{xz}\sigma_{zx} + 2\sigma_{yz}\sigma_{zy})\right] \\[2mm]
&=\; \frac{1}{2}\left[\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xx}\sigma_{yy} + 2\sigma_{xx}\sigma_{zz} + 2\sigma_{yy}\sigma_{zz} - (\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xy}\sigma_{yx} + 2\sigma_{xz}\sigma_{zx} + 2\sigma_{yz}\sigma \right. \\[2mm]
&=\; \frac{1}{2}\left[2\sigma_{xx}\sigma_{yy} + 2\sigma_{xx}\sigma_{zz} + 2\sigma_{yy}\sigma_{zz} - (2\sigma_{xy}\sigma_{yx} + 2\sigma_{xz}\sigma_{zx} + 2\sigma_{yz}\sigma_{zy})\right] \\[2mm]
&=\; \sigma_{xx}\sigma_{yy} + \sigma_{xx}\sigma_{zz} + \sigma_{yy}\sigma_{zz} - \sigma_{xy}\sigma_{yx} - \sigma_{xz}\sigma_{zx} - \sigma_{yz}\sigma_{zy} \\[2mm]
&=\; \sigma_{xx}\sigma_{yy} + \sigma_{yy}\sigma_{zz} + \sigma_{xx}\sigma_{zz} - \sigma_{xy}^2 - \sigma_{xz}^2 - \sigma_{yz}^2
\end{aligned}
$$

Let us now express the second invariant of the deviatoric stress tensor as a function of the invariants of the full stress tensor (just to be sure I have carried this out twice in what follows):

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\sum_{ij}\tau_{ij}\tau_{ji} \\
&= \frac{1}{2}\sum_{ij}\left(\sigma_{ij} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right)\left(\sigma_{ij} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right) \\
&= \frac{1}{2}\sum_{ij}\left[\sigma_{ij}\sigma_{ij} + \sigma_{ij}\left(-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right) + \sigma_{ij}\left(-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right) + \left(-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right)\left(-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right)\right] \\
&= \frac{1}{2}\sum_{ij}\left[\sigma_{ij}\sigma_{ij} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\sigma_{ij}\delta_{ij} + \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\delta_{ij}\right] \\
&= \underbrace{\frac{1}{2}\sum_{ij}\sigma_{ij}\sigma_{ij}}_{\mathcal{I}_2(\boldsymbol{\sigma})} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\sum_{ij}\sigma_{ij}\delta_{ij}}_{\mathcal{I}_1(\boldsymbol{\sigma})} + \frac{1}{18}\mathcal{I}_1(\boldsymbol{\sigma})^2\underbrace{\sum_{ij}\delta_{ij}}_{3} \\
&= \mathcal{I}_2(\boldsymbol{\sigma}) - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 \\
&= -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_2(\boldsymbol{\tau}) &= \frac{1}{2}\boldsymbol{\tau}:\boldsymbol{\tau} \\
&= \frac{1}{2}(\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2) + \tau_{xy}^2 + \tau_{xz}^2 + \tau_{yz}^2 \\
&= \frac{1}{2}\left((\sigma_{xx} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))^2 + (\sigma_{yy} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))^2 + (\sigma_{zz} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))^2\right) + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= \frac{1}{2}\left(\sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1(\boldsymbol{\sigma}) + \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1(\boldsymbol{\sigma}) + \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1(\boldsymbol{\sigma}) + \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\right) + \\
&= \frac{1}{2}\left(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{2}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})\mathcal{I}_1(\boldsymbol{\sigma}) + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})^2\right) + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= \frac{1}{2}\left(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})^2\right) + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= \frac{1}{2}\left(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})^2\right) + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \frac{1}{2}\left(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2\right) + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\
&= -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma})
\end{aligned}
$$

So there is no doubt:

$$
\boxed{\mathcal{I}_2(\boldsymbol{\tau}) = -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma})}
$$

Note that this relationship is often found is a very confusing form where moment invariants $\mathcal{K}_{1,2,3}$ are used instead of principal invariants $\mathcal{I}_{1,2,3}$ (although the letter I is used!). We have established that $\mathcal{I}_2(\boldsymbol{\sigma}) = \frac{1}{2}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma})$ with $\mathcal{K}_1(\boldsymbol{\sigma}) = \mathcal{I}_1(\boldsymbol{\sigma})$ so that

$$
\mathcal{I}_2(\boldsymbol{\tau}) = -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma}) = -\frac{1}{6}\mathcal{K}_1(\boldsymbol{\sigma})^2 + \frac{1}{2}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma}) = \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma})
$$

that is:

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma})$$

If one replaces $\mathcal{K}$'s by $\mathcal{I}$'s then one finds the formula in the literature [1] [2].

---

[1] https://www.pantelisliolios.com/deviatoric-stress-and-invariants/
[2] https://en.wikipedia.org/wiki/Cauchy_stress_tensor

third invariants

Let us now look at the third invariant:

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{1}{3}\sum_{ijk}\tau_{ij}\tau_{jk}\tau_{ki} \\
&= \frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^3 + \tau_{zz}^3) + \tau_{xx}(\tau_{xy}^2 + \tau_{xz}^2) + \tau_{yy}(\tau_{xy}^2 + \tau_{yz}^2) + \tau_{zz}(\tau_{xz}^2 + \tau_{yz}^2) + 2\tau_{xy}\tau_{xz}\tau_{yz}
\end{aligned}
$$

$$
\begin{aligned}
I_3(\boldsymbol{\tau}) &= \det(\boldsymbol{\tau}) \\
&= \tau_{xx}(\tau_{yy}\tau_{zz} - \tau_{yz}^2) - \tau_{yx}(\tau_{xy}\tau_{zz} - \tau_{zy}\tau_{xz}) + \tau_{zx}(\tau_{xy}\tau_{yz} - \tau_{yy}\tau_{xz}) \\
&= \tau_{xx}\tau_{yy}\tau_{zz} - \tau_{xx}\tau_{yz}^2 - \tau_{zz}\tau_{xy}^2 + \tau_{xy}\tau_{yz}\tau_{yz} + \tau_{xy}\tau_{yz}\tau_{yz} - \tau_{yy}\tau_{xz}^2 \\
&= \tau_{xx}\tau_{yy}\tau_{zz} - \tau_{xx}\tau_{yz}^2 - \tau_{zz}\tau_{xy}^2 + 2\tau_{xy}\tau_{yz}\tau_{yz} - \tau_{yy}\tau_{xz}^2 \\
&= \tau_{xx}\tau_{yy}\tau_{zz} - (-\tau_{yy} - \tau_{zz})\tau_{yz}^2 - (-\tau_{xx} - \tau_{yy})\tau_{xy}^2 + 2\tau_{xy}\tau_{yz}\tau_{yz} - (-\tau_{xx} - \tau_{zz})\tau_{xz}^2 \\
&= \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{xx}(\tau_{xy}^2 + \tau_{xz}^2) + \tau_{yy}(\tau_{xy}^2 + \tau_{yz}^2) + \tau_{zz}(\tau_{xz}^2 + \tau_{yz}^2) + 2\tau_{xy}\tau_{xz}\tau_{yz}
\end{aligned}
$$

The first term is still different than $\frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^3 + \tau_{zz}^3)$... or is it? Let us have a go using the fact that $\boldsymbol{\tau}$ is deviatoric:

$$
\begin{aligned}
\tau_{xx}\tau_{yy}\tau_{zz} &= \tau_{xx}(-\tau_{xx} - \tau_{zz})(-\tau_{xx} - \tau_{yy}) \\
&= \tau_{xx}(\tau_{xx}^2 + \tau_{xx}\tau_{yy} + \tau_{xx}\tau_{zz} + \tau_{yy}\tau_{zz}) \\
&= \tau_{xx}^3 + \tau_{xx}^2\tau_{yy} + \tau_{xx}^2\tau_{zz} + \tau_{xx}\tau_{yy}\tau_{zz} \\
\tau_{xx}\tau_{yy}\tau_{zz} &= (-\tau_{yy} - \tau_{zz})\tau_{yy}(-\tau_{xx} - \tau_{yy}) \\
&= \tau_{yy}(\tau_{xx}\tau_{yy} + \tau_{yy}^2 + \tau_{xx}\tau_{zz} + \tau_{yy}\tau_{zz}) \\
&= \tau_{xx}\tau_{yy}^2 + \tau_{yy}^3 + \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{yy}^2\tau_{zz} \\
\tau_{xx}\tau_{yy}\tau_{zz} &= (-\tau_{yy} - \tau_{zz})(-\tau_{xx} - \tau_{zz})\tau_{zz} \\
&= \tau_{zz}(\tau_{xx}\tau_{yy} + \tau_{yy}\tau_{zz} + \tau_{xx}\tau_{zz} + \tau_{zz}^2) \\
&= \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{yy}\tau_{zz}^2 + \tau_{xx}\tau_{zz}^2 + \tau_{zz}^3
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \quad \tau_{xx}\tau_{yy}\tau_{zz} &= \frac{1}{3}(\tau_{xx}\tau_{yy}\tau_{zz} + \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{xx}\tau_{yy}\tau_{zz}) \\
&= \frac{1}{3}(\tau_{xx}^3 + \tau_{xx}^2\tau_{yy} + \tau_{xx}^2\tau_{zz} + \tau_{xx}\tau_{yy}\tau_{zz} \\
&\quad + \tau_{xx}\tau_{yy}^2 + \tau_{yy}^3 + \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{yy}^2\tau_{zz} \\
&\quad + \tau_{xx}\tau_{yy}\tau_{zz} + \tau_{yy}\tau_{zz}^2 + \tau_{xx}\tau_{zz}^2 + \tau_{zz}^3) \\
&= \frac{1}{3}[\tau_{xx}^3 + \tau_{yy}^3 + \tau_{zz}^3 + \tau_{xx}\tau_{yy}\underbrace{(\tau_{xx} + \tau_{yy} + \tau_{zz})}_{=0} + \tau_{xx}\tau_{zz}\underbrace{(\tau_{xx} + \tau_{yy} + \tau_{zz})}_{=0} + \tau_{yy}\tau_{zz}\underbrace{(\tau_{xx} + \tau_{yy} +}_{=0} \\
&= \frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^3 + \tau_{zz}^3)
\end{aligned}
$$

Let us now turn to $I_3(\boldsymbol{\tau}) = \frac{1}{3}\text{tr}[\boldsymbol{\tau} \cdot \boldsymbol{\tau} \cdot \boldsymbol{\tau}]$. Assuming the tensor $\boldsymbol{\tau}$ to be symmetric then

$$
\boldsymbol{\tau} = \begin{pmatrix} a & d & e \\ d & b & f \\ e & f & c \end{pmatrix}
$$

then, thanks to `https://www.wolframalpha.com/` I find that

$$
\boldsymbol{\tau}\cdot\boldsymbol{\tau}\cdot\boldsymbol{\tau} = \begin{pmatrix} a(a^2 + d^2 + e^2) + d(ad + bd + ef) + e(ae + ce + df) & d(a^2 + d^2 + e^2) + b(ad + bd + ef) + f(ae + ce + df) & e(a^2 + d^2 + e^2) + f(ad + bd + ef) + c(ae + ce + \\ a(ad + bd + ef) + d(b^2 + d^2 + f^2) + e(bf + cf + de) & d(ad + bd + ef) + b(b^2 + d^2 + f^2) + f(bf + cf + de) & e(ad + bd + ef) + f(b^2 + d^2 + f^2) + c(bf + cf + \\ a(ae + ce + df) + d(bf + cf + de) + e(c^2 + e^2 + f^2) & d(ae + ce + df) + b(bf + cf + de) + f(c^2 + e^2 + f^2) & e(ae + ce + df) + f(bf + cf + de) + c(c^2 + e^2 + \end{pmatrix}
$$

and then

$$\frac{1}{3}\mathrm{tr}[\boldsymbol{\tau}\cdot\boldsymbol{\tau}\cdot\boldsymbol{\tau}] = \frac{1}{3}(a^3+b^3+c^3)+c(e^2+f^2)+a(d^2+e^2)+2def+b(d^2+f^2)$$

$$= \frac{1}{3}(\tau_{xx}^3+\tau_{yy}^3+\tau_{zz}^3)+\tau_{xx}(\tau_{xy}^2+\tau_{xz}^2)+\tau_{yy}(\tau_{xy}^2+\tau_{yz}^2)+\tau_{zz}(\tau_{xz}^2+\tau_{yz}^2)+2\tau_{xy}\tau_{xz}\tau_{yz}$$

Let us now express the third invariant of the deviatoric stress tensor as a function of the invariants of the full stress tensor (just to be sure I have carried this out twice in what follows):

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{1}{3}\sum_{ijk}\tau_{ij}\tau_{jk}\tau_{ki}\\[4pt]
&= \frac{1}{3}\sum_{ijk}\left(\sigma_{ij}-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\right)\tau_{jk}\tau_{ki}\\[4pt]
&= \frac{1}{3}\sum_{ijk}\left[\sigma_{ij}\tau_{jk}\tau_{ki}-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\tau_{jk}\tau_{ki}\right]\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\tau_{jk}\tau_{ki}-\frac{1}{9}\sum_{ijk}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ij}\tau_{jk}\tau_{ki}\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\tau_{jk}\tau_{ki}-\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})\sum_{ik}\tau_{ik}\tau_{ki}\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\tau_{jk}\tau_{ki}-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\frac{1}{2}\sum_{ik}\tau_{ik}\tau_{ki}}_{\mathcal{I}_2(\boldsymbol{\tau})}\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\tau_{jk}\tau_{ki}-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\left(\sigma_{jk}-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{jk}\right)\left(\sigma_{ki}-\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ki}\right)-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})\\[4pt]
&= \frac{1}{3}\sum_{ijk}\left(\sigma_{ij}\sigma_{jk}\sigma_{ki}-\sigma_{ij}\sigma_{jk}\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ki}-\sigma_{ij}\sigma_{ki}\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{jk}+\sigma_{ij}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\delta_{jk}\delta_{ki}\right)-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})\\[4pt]
&= \frac{1}{3}\sum_{ijk}\sigma_{ij}\sigma_{jk}\sigma_{ki}-\frac{1}{3}\sum_{ijk}\sigma_{ij}\sigma_{jk}\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{ki}-\frac{1}{3}\sum_{ijk}\sigma_{ij}\sigma_{ki}\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\delta_{jk}+\frac{1}{3}\sum_{ijk}\sigma_{ij}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\delta_{jk}\delta_{ki}-\frac{2}{9}\mathcal{I}_1(\\[4pt]
&= \underbrace{\frac{1}{3}\sum_{ijk}\sigma_{ij}\sigma_{jk}\sigma_{ki}}_{\mathcal{I}_3(\boldsymbol{\sigma})}-\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})\sum_{ij}\sigma_{ij}\sigma_{ji}-\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})\sum_{ij}\sigma_{ij}\sigma_{ji}+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^2\underbrace{\sum_{ijk}\sigma_{ij}\delta_{jk}\delta_{ki}}_{\mathcal{I}_1(\boldsymbol{\sigma})}-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})\\[4pt]
&= \mathcal{I}_3(\boldsymbol{\sigma})-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\frac{1}{2}\sum_{ij}\sigma_{ij}\sigma_{ji}}_{\mathcal{I}_2(\boldsymbol{\sigma})}-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\frac{1}{2}\sum_{ij}\sigma_{ij}\sigma_{ji}}_{\mathcal{I}_2(\boldsymbol{\sigma})}+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})\\[4pt]
&= \mathcal{I}_3(\boldsymbol{\sigma})-\frac{4}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma})+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\tau})
\end{aligned}
$$

Then we use $\mathcal{I}_2(\boldsymbol{\tau})=-\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2+\mathcal{I}_2(\boldsymbol{\sigma})$ so

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \mathcal{I}_3(\boldsymbol{\sigma})-\frac{4}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\left(-\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2+\mathcal{I}_2(\boldsymbol{\sigma})\right)\\[4pt]
&= \mathcal{I}_3(\boldsymbol{\sigma})-\frac{4}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3+\frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3-\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma})\\[4pt]
&= \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3-\frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma})+\mathcal{I}_3(\boldsymbol{\sigma})
\end{aligned}
$$

We start this time from

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^2 + \tau_{zz}^3) + \tau_{xx}(\tau_{xy}^2 + \tau_{xz}^2) + \tau_{yy}(\tau_{xy}^2 + \tau_{yz}^2) + \tau_{zz}(\tau_{xz}^2 + \tau_{yz}^2) + 2\tau_{xy}\tau_{yz}\tau_{yz}$$

We have

$$\tau_{xx}^3 = \left(\sigma_{xx} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\right)^3$$

$$= \sigma_{xx}^3 - 3\sigma_{xx}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{xx}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3$$

$$\tau_{yy}^3 = \left(\sigma_{yy} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\right)^3$$

$$= \sigma_{yy}^3 - 3\sigma_{yy}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{yy}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3$$

$$\tau_{zz}^3 = \left(\sigma_{zz} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\right)^3$$

$$= \sigma_{zz}^3 - 3\sigma_{zz}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{zz}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3$$

Then

$$\begin{aligned}
\tau_{xx}^3 + \tau_{yy}^2 + \tau_{zz}^3 &= \sigma_{xx}^3 - 3\sigma_{xx}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{xx}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&+ \sigma_{yy}^3 - 3\sigma_{yy}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{yy}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&+ \sigma_{zz}^3 - 3\sigma_{zz}^2\frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}) + 3\sigma_{zz}\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&= \sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3 - \mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2) + \frac{1}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&= \sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3 - \mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2) + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&= \sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3 - \mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2) + \frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^3
\end{aligned}$$

$$\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{1}{3}(\tau_{xx}^3 + \tau_{yy}^2 + \tau_{zz}^3) + \tau_{xx}(\tau_{xy}^2 + \tau_{xz}^2) + \tau_{yy}(\tau_{xy}^2 + \tau_{yz}^2) + \tau_{zz}(\tau_{xz}^2 + \tau_{yz}^2) + 2\tau_{xy}\tau_{yz}\tau_{yz} \\
&= \frac{1}{3}(\sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3) - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2) + \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&\quad + (\sigma_{xx} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))(\sigma_{xy}^2 + \sigma_{xz}^2) + (\sigma_{yy} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))(\sigma_{xy}^2 + \sigma_{yz}^2) + (\sigma_{zz} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma}))(\sigma_{xz}^2 + \sigma_{yz}^2) + 2\sigma_{xy}\sigma_{yz}\sigma \\
&= \frac{1}{3}(\sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3) - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{xy}^2 + \sigma_{yz}^2 + \sigma_{xz}^2 + \sigma_{yz}^2) + \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&\quad + \sigma_{xx}(\sigma_{xy}^2 + \sigma_{xz}^2) + \sigma_{yy}(\sigma_{xy}^2 + \sigma_{yz}^2) + \sigma_{zz}(\sigma_{xz}^2 + \sigma_{yz}^2) + 2\sigma_{xy}\sigma_{yz}\sigma_{yz} \\
&= \frac{1}{3}(\sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3) - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xy}^2 + 2\sigma_{xz}^2 + 2\sigma_{yz}^2) + \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&\quad + \sigma_{xx}(\sigma_{xy}^2 + \sigma_{xz}^2) + \sigma_{yy}(\sigma_{xy}^2 + \sigma_{yz}^2) + \sigma_{zz}(\sigma_{xz}^2 + \sigma_{yz}^2) + 2\sigma_{xy}\sigma_{yz}\sigma_{yz} \\
&= \frac{1}{3}(\sigma_{xx}^3 + \sigma_{yy}^3 + \sigma_{zz}^3) - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\frac{1}{2}(\sigma_{xx}^2 + \sigma_{yy}^2 + \sigma_{zz}^2 + 2\sigma_{xy}^2 + 2\sigma_{xz}^2 + 2\sigma_{yz}^2)}_{\mathcal{I}_2(\boldsymbol{\sigma})} + \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 \\
&\quad + \sigma_{xx}(\sigma_{xy}^2 + \sigma_{xz}^2) + \sigma_{yy}(\sigma_{xy}^2 + \sigma_{yz}^2) + \sigma_{zz}(\sigma_{xz}^2 + \sigma_{yz}^2) + 2\sigma_{xy}\sigma_{yz}\sigma_{yz} \\
&= \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma}) + \mathcal{I}_3(\boldsymbol{\sigma})
\end{aligned}$$

Then, without doubt

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma}) + \mathcal{I}_3(\boldsymbol{\sigma})$$

Let us now rewrite this relationship as a function of the principal invariants using the following relationships:

$$
\begin{aligned}
\mathcal{I}_1(\boldsymbol{\sigma}) &= \mathcal{K}_1(\boldsymbol{\sigma}) \\
\mathcal{I}_2(\boldsymbol{\sigma}) &= \frac{1}{2}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma}) \\
\mathcal{I}_3(\boldsymbol{\sigma}) &= \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_3(\boldsymbol{\tau}) &= \frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma}) + \mathcal{I}_3(\boldsymbol{\sigma}) \\
&= \frac{2}{27}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{K}_1(\boldsymbol{\sigma})\left(\frac{1}{2}\mathcal{K}_1(\boldsymbol{\sigma})^2 - \mathcal{K}_2(\boldsymbol{\sigma})\right) + \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma}) \\
&= \frac{2}{27}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^3 + \frac{2}{3}\mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma}) \\
&= \frac{2}{27}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma}) \quad\quad\quad\text{(Q.5)}
\end{aligned}
$$

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{2}{27}\mathcal{K}_1(\boldsymbol{\sigma})^3 - \frac{1}{3}\mathcal{K}_1(\boldsymbol{\sigma})\mathcal{K}_2(\boldsymbol{\sigma}) + \mathcal{K}_3(\boldsymbol{\sigma}) \quad\quad\quad\text{(Q.6)}$$

If one replaces the $\mathcal{K}$'s by I's then one finds the formula in the literature [3] [4].

$$
\begin{aligned}
\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} &= \frac{\partial}{\partial \boldsymbol{\sigma}}\left(\frac{2}{27}\mathcal{I}_1(\boldsymbol{\sigma})^3 - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathcal{I}_2(\boldsymbol{\sigma}) + \mathcal{I}_3(\boldsymbol{\sigma})\right) \\
&= \frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 \underbrace{\frac{\partial\mathcal{I}_1(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}}}_{\mathbf{1}} - \frac{2}{3}\underbrace{\frac{\partial\mathcal{I}_1(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}}}_{\mathbf{1}}\mathcal{I}_2(\boldsymbol{\sigma}) - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\underbrace{\frac{\partial\mathcal{I}_2(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}}}_{\boldsymbol{\sigma}} + \underbrace{\frac{\partial\mathcal{I}_3(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}}}_{\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}} \\
&= \frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\mathbf{1} - \frac{2}{3}\mathbf{1}\mathcal{I}_2(\boldsymbol{\sigma}) - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\sigma} + \boldsymbol{\sigma}\cdot\boldsymbol{\sigma} \\
&= \left(\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\sigma})\right)\mathbf{1} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\sigma} + \boldsymbol{\sigma}\cdot\boldsymbol{\sigma} \quad\quad\text{(Q.7)}
\end{aligned}
$$

Using $\mathcal{I}_2(\boldsymbol{\tau}) = -\frac{1}{6}\mathcal{I}_1(\boldsymbol{\sigma})^2 + \mathcal{I}_2(\boldsymbol{\sigma})$:

$$
\begin{aligned}
\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} &= \left(\frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau}) - \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\right)\mathbf{1} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\sigma} + \boldsymbol{\sigma}\cdot\boldsymbol{\sigma} \\
&= \left(\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\right)\mathbf{1} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\sigma} + \boldsymbol{\sigma}\cdot\boldsymbol{\sigma} \\
&= \left(\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\right)\mathbf{1} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\left(\boldsymbol{\tau} + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathbf{1}\right) + \left(\boldsymbol{\tau} + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathbf{1}\right)\cdot\left(\boldsymbol{\tau} + \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathbf{1}\right) \\
&= \left(\frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2 - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\right)\mathbf{1} - \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\tau} - \frac{2}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\mathbf{1} + \boldsymbol{\tau}\cdot\boldsymbol{\tau} + \frac{2}{3}\mathcal{I}_1(\boldsymbol{\sigma})\boldsymbol{\tau} + \frac{1}{9}\mathcal{I}_1(\boldsymbol{\sigma})^2\mathbf{1} \\
&= \boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1} \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\text{(Q.8)}
\end{aligned}
$$

---

[3] https://www.pantelisliolios.com/deviatoric-stress-and-invariants/
[4] https://en.wikipedia.org/wiki/Cauchy_stress_tensor

which is the so-called Hill tensor[5].

Note that this tensor is deviatoric:

$$\text{tr}\left[\boldsymbol{\tau}\cdot\boldsymbol{\tau} - \frac{2}{3}\mathcal{I}_2(\boldsymbol{\tau})\mathbf{1}\right] = \text{tr}[\boldsymbol{\tau}\cdot\boldsymbol{\tau}] - 2\mathcal{I}_2(\boldsymbol{\tau}) = 2\mathcal{I}_2(\boldsymbol{\tau}) - 2\mathcal{I}_2(\boldsymbol{\tau}) = 0$$

---

[5]https://en.wikipedia.org/wiki/Lode_coordinates

Derivatives

The derivatives of the invariants with respect to the stress tensor are tensors given as follows:

$$
\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \begin{pmatrix} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{xx}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{xy}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{xz}} \\ \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{yx}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{yy}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{yz}} \\ \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{zx}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{zy}} & \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \sigma_{zz}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{1} \tag{Q.9}
$$

$$
\frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \begin{pmatrix} \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{xx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{xy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{xz}} \\ \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{yx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{yy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{yz}} \\ \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{zx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{zy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\sigma})}{\partial \sigma_{zz}} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2\sigma_{xx} & 2\sigma_{xy} & 2\sigma_{xz} \\ 2\sigma_{yx} & 2\sigma_{yy} & 2\sigma_{yz} \\ 2\sigma_{zx} & 2\sigma_{zy} & 2\sigma_{zz} \end{pmatrix} = \boldsymbol{\sigma} \tag{Q.10}
$$

$$
\frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \begin{pmatrix} \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{xx}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{xy}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{xz}} \\ \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{yx}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{yy}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{yz}} \\ \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{zx}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{zy}} & \frac{\partial \mathcal{I}_3(\boldsymbol{\sigma})}{\partial \sigma_{zz}} \end{pmatrix}
$$

$$
= \begin{pmatrix} \sigma_{xx}^2 + \sigma_{xy}^2 + \sigma_{xz}^2 & \sigma_{xx}\sigma_{xy} + \sigma_{yy}\sigma_{xy} + \sigma_{xz}\sigma_{yz} & \sigma_{xx}\sigma_{xz} + \sigma_{zz}\sigma_{xz} + \sigma_{xy}\sigma_{yz} \\ ... & \sigma_{yy}^2 + \sigma_{xy}^2 + \sigma_{yz}^2 & ... \\ ... & ... & ...\sigma_{zz}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \end{pmatrix}
$$

$$
= \boldsymbol{\sigma} \cdot \boldsymbol{\sigma} \tag{Q.11}
$$

where we have used the generic form of the second and third invariants, i.e. not assuming the tensors to be symmetric so that (for example) σ_{xz} and σ_{zx} are distinct quantities.

The Lodé angle $\theta_{\mathrm{L}}(\boldsymbol{\tau})$ is actually a function of $\mathcal{I}_2(\boldsymbol{\tau})$ and $\mathcal{I}_3(\boldsymbol{\tau})$ as follows:

$$
\sin 3\theta_{\mathrm{L}}(\boldsymbol{\tau}) = -\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}}
$$

Since this quantity unambiguously depends on the deviatoric stress tensor, I will omit the '$(\boldsymbol{\tau})$' dependency in what follows. Then

$$
\frac{\partial}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \sin 3\theta_{\mathrm{L}} = 3\cos 3\theta_{\mathrm{L}}(\boldsymbol{\tau}) \frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_2(\boldsymbol{\tau})}
$$

$$
\frac{\partial}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \sin 3\theta_{\mathrm{L}} = 3\cos 3\theta_{\mathrm{L}}(\boldsymbol{\tau}) \frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_3(\boldsymbol{\tau})}
$$

so that

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_2(\boldsymbol{\tau})} = \frac{1}{3\cos 3\theta_{\mathrm{L}}} \frac{\partial}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \sin 3\theta_{\mathrm{L}}$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \frac{\partial}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \left(-\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right)$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \left(-\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right) \left(-\frac{3}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \right)$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \sin 3\theta_{\mathrm{L}} \left(-\frac{3}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \right)$$

$$= -\frac{1}{2} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})}$$

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_3(\boldsymbol{\tau})} = \frac{1}{3\cos 3\theta_{\mathrm{L}}} \frac{\partial}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \sin 3\theta_{\mathrm{L}}$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \frac{\partial}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \left(-\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right)$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \left(-\frac{3\sqrt{3}}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right)$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \left(-\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right) \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})}$$

$$= \frac{1}{3\cos 3\theta_{\mathrm{L}}} \sin 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})}$$

$$= \frac{1}{3} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})}$$

We have just established the useful relationships

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_2(\boldsymbol{\tau})} = -\frac{1}{2} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \tag{Q.12}$$

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \mathcal{I}_3(\boldsymbol{\tau})} = \frac{1}{3} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})} \tag{Q.13}$$

and in the end we can write

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \boldsymbol{\sigma}} = \frac{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{\partial \theta_{\mathrm{L}}(\boldsymbol{\tau})}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}}$$

$$= \left(-\frac{1}{2} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \right) \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \left(\frac{1}{3} \tan 3\theta_{\mathrm{L}} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})} \right) \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}}$$

$$= \tan 3\theta_{\mathrm{L}} \left[-\frac{1}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{3} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right]$$

$$= \frac{\sin 3\theta_{\mathrm{L}}}{\cos 3\theta_{\mathrm{L}}} \left[-\frac{1}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{3} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right]$$

$$= \frac{1}{\cos 3\theta_{\mathrm{L}}} \left(-\frac{3\sqrt{3}}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \right) \left[-\frac{1}{2} \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{3} \frac{1}{\mathcal{I}_3(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right]$$

$$= -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \left[-\frac{3}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{5/2}} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right] \tag{Q.14}$$

i.e.

$$\frac{\partial \theta_{\mathrm{L}}}{\partial \boldsymbol{\sigma}} = \quad = \quad -\frac{\sqrt{3}}{2\cos 3\theta_{\mathrm{L}}} \left[-\frac{3}{2} \frac{\mathcal{I}_3(\boldsymbol{\tau})}{\mathcal{I}_2(\boldsymbol{\tau})^{5/2}} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} + \frac{1}{\mathcal{I}_2(\boldsymbol{\tau})^{3/2}} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} \right] \qquad \text{(Q.15)}$$

which is Eq. (7.68) of Owen & Hinton:

$$\frac{\partial \theta}{\partial \boldsymbol{\sigma}} = \frac{-\sqrt{3}}{2\cos 3\theta} \left[\frac{1}{(J_2')^{3/2}} \frac{\partial J_3}{\partial \boldsymbol{\sigma}} - \frac{3 J_3}{(J_2')^2} \frac{\partial (J_2')^{1/2}}{\partial \boldsymbol{\sigma}} \right]. \qquad \text{(7.68)}$$

Taken from Owen and Hinton [1026]

Appendix R

The Γ tensor in plasticity

WARNING: this is not finished.

Let us start from

$$\vec{\dot{\varepsilon}} = \mathbf{\Gamma}(\vec{\sigma}) \cdot \vec{\sigma}$$

Note that we will later need $\mathbf{\Gamma}^{-1}$ which begs the question of it being invertible... In Zienkiewicz [1532], the author states: "In many forms of the visco-plastic law the relationship $\vec{\dot{\varepsilon}} = \mathbf{\Gamma}(\vec{\sigma}) \cdot \vec{\sigma}$ is such that no volumetric strain rate exists i.e. the material is incompressible. Now $\mathbf{\Gamma}$ does not posses an inverse".

R.0.1 Computing the Γ matrix

Let us first establish that we can write quite generally in the isotropic case

$$\frac{\partial Q}{\partial \vec{\sigma}} = \mathbf{\Gamma}_0 \cdot \vec{\sigma}$$

By applying the chain rule we can write

$$
\begin{aligned}
\frac{\partial Q}{\partial \vec{\sigma}} &= \frac{\partial Q}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \vec{\sigma}} + \frac{\partial Q}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \vec{\sigma}} + \frac{\partial Q}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \vec{\sigma}} \\
&= \left(\frac{\partial Q}{\partial \mathcal{I}_1(\boldsymbol{\sigma})} \boldsymbol{M}_1(\boldsymbol{\sigma}) + \frac{\partial Q}{\partial \mathcal{I}_2(\boldsymbol{\tau})} \boldsymbol{M}_2(\boldsymbol{\sigma}) + \frac{\partial Q}{\partial \mathcal{I}_3(\boldsymbol{\tau})} \boldsymbol{M}_3(\boldsymbol{\sigma}) \right) \cdot \vec{\sigma}
\end{aligned}
\tag{R.1}
$$

All we have to do now is to compute the three symmetric matrices $\boldsymbol{M}_{1,2,3}(\boldsymbol{\sigma})$ which independent of F or Q.

Computing matrix M_1

$$\frac{\partial \mathcal{I}_1(\boldsymbol{\sigma})}{\partial \vec{\sigma}} = \frac{\partial}{\partial \vec{\sigma}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})$$

$$= \begin{pmatrix} \frac{\partial}{\partial \sigma_{xx}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \\ \frac{\partial}{\partial \sigma_{yy}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \\ \frac{\partial}{\partial \sigma_{zz}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \\ \frac{\partial}{\partial \sigma_{xy}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \\ \frac{\partial}{\partial \sigma_{xz}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \\ \frac{\partial}{\partial \sigma_{yz}}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$= \frac{1}{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}} \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \vec{\sigma}$$

$$= \frac{1}{\mathcal{I}_1(\boldsymbol{\sigma})} \underbrace{\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{\boldsymbol{M_1(\sigma)}} \cdot \vec{\sigma} \qquad \text{(R.2)}$$

Computing matrix M_2

We start from

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6}\left[(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{xx} - \sigma_{zz})^2\right] + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \qquad \text{(R.3)}$$

Then

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \vec{\sigma}} = \begin{pmatrix} \frac{\partial}{\partial \sigma_{xx}} \mathcal{I}_2(\boldsymbol{\tau}) \\ \frac{\partial}{\partial \sigma_{yy}} \mathcal{I}_2(\boldsymbol{\tau}) \\ \frac{\partial}{\partial \sigma_{zz}} \mathcal{I}_2(\boldsymbol{\tau}) \\ \frac{\partial}{\partial \sigma_{xy}} \mathcal{I}_2(\boldsymbol{\tau}) \\ \frac{\partial}{\partial \sigma_{xz}} \mathcal{I}_2(\boldsymbol{\tau}) \\ \frac{\partial}{\partial \sigma_{yz}} \mathcal{I}_2(\boldsymbol{\tau}) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{6}(2(\sigma_{xx} - \sigma_{yy}) + 2(\sigma_{xx} - \sigma_{zz})) \\ \frac{1}{6}(-2(\sigma_{xx} - \sigma_{yy}) + 2(\sigma_{yy} - \sigma_{zz})) \\ \frac{1}{6}(-2(\sigma_{yy} - \sigma_{zz}) - 2(\sigma_{xx} - \sigma_{zz})) \\ 2\sigma_{xy} \\ 2\sigma_{xz} \\ 2\sigma_{yz} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{2}{3}\sigma_{xx} - \frac{1}{3}\sigma_{yy} - \frac{1}{3}\sigma_{zz} \\ -\frac{2}{3}\sigma_{xx} + \frac{4}{3}\sigma_{yy} - \frac{2}{3}\sigma_{zz} \\ -\frac{2}{3}\sigma_{xx} - \frac{2}{3}\sigma_{yy} + \frac{4}{3}\sigma_{zz} \\ 2\sigma_{xy} \\ 2\sigma_{xz} \\ 2\sigma_{yz} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 2/3 & -1/3 & -1/3 & 0 & 0 & 0 \\ -1/3 & 2/3 & -1/3 & 0 & 0 & 0 \\ -1/3 & -1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}}_{\boldsymbol{M}_2(\boldsymbol{\sigma})} \cdot \vec{\sigma} \tag{R.4}$$

This is the same matrix as in Eq. (13.11) in Zienkiewicz [1532] (1975).

Another look at it using tensors: we start this time from

$$\mathcal{I}_2(\boldsymbol{\tau}) = \frac{1}{6}\left[(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{xx} - \sigma_{zz})^2\right] + \frac{1}{2}(\sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 + \sigma_{yx}^2 + \sigma_{zx}^2 + \sigma_{zy}^2) \tag{R.5}$$

Then

$$\frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \boldsymbol{\sigma}} = \begin{pmatrix} \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{xz}} \\ \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{yz}} \\ \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{zx}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{zy}} & \frac{\partial \mathcal{I}_2(\boldsymbol{\tau})}{\partial \sigma_{zz}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{2}{3}\sigma_{xx} - \frac{1}{3}\sigma_{yy} - \frac{1}{3}\sigma_{zz} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & -\frac{2}{3}\sigma_{xx} + \frac{4}{3}\sigma_{yy} - \frac{2}{3}\sigma_{zz} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & -\frac{2}{3}\sigma_{xx} - \frac{2}{3}\sigma_{yy} + \frac{4}{3}\sigma_{zz} \end{pmatrix}$$

$$= \boldsymbol{\sigma} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathbf{1}$$

$$= \boldsymbol{\tau} \tag{R.6}$$

Computing matrix M_3

The third invariant proves to be the most annoying:

$$\mathcal{I}_3(\boldsymbol{\tau}) = \frac{1}{3}\sum_{i,j,k}\tau_{ij}\tau_{jk}\tau_{ki} \tag{R.7}$$

$$= \frac{1}{3}\tau_{xx}(\tau_{xx}^2 + 3\tau_{xy}^2 + 3\tau_{xz}^2)$$

$$+ \frac{1}{3}\tau_{yy}(3\tau_{xy}^2 + \tau_{yy}^2 + 3\tau_{yz}^2)$$

$$+ \frac{1}{3}\tau_{zz}(3\tau_{xz}^2 + 3\tau_{yz}^2 + \tau_{zz}^2)$$

$$+ 2\tau_{xy}\tau_{xz}\tau_{yz} \tag{R.8}$$

Then

$$\frac{\partial\mathcal{I}_3(\boldsymbol{\tau})}{\partial\vec{\sigma}} = \begin{pmatrix} \frac{\partial}{\partial\sigma_{xx}}\mathcal{I}_3(\boldsymbol{\tau}) \\ \frac{\partial}{\partial\sigma_{yy}}\mathcal{I}_3(\boldsymbol{\tau}) \\ \frac{\partial}{\partial\sigma_{zz}}\mathcal{I}_3(\boldsymbol{\tau}) \\ \frac{\partial}{\partial\sigma_{xy}}\mathcal{I}_3(\boldsymbol{\tau}) \\ \frac{\partial}{\partial\sigma_{xz}}\mathcal{I}_3(\boldsymbol{\tau}) \\ \frac{\partial}{\partial\sigma_{yz}}\mathcal{I}_3(\boldsymbol{\tau}) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial}{\partial\tau_{xx}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{xx}}{\partial\sigma_{xx}} \\ \frac{\partial}{\partial\tau_{yy}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{yy}}{\partial\sigma_{yy}} \\ \frac{\partial}{\partial\tau_{zz}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{zz}}{\partial\sigma_{zz}} \\ \frac{\partial}{\partial\tau_{xy}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{xy}}{\partial\sigma_{xy}} \\ \frac{\partial}{\partial\tau_{xz}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{xz}}{\partial\sigma_{xz}} \\ \frac{\partial}{\partial\tau_{yz}}\mathcal{I}_3(\boldsymbol{\tau})\frac{\partial\tau_{yz}}{\partial\sigma_{yz}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial}{\partial\tau_{xx}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ \frac{\partial}{\partial\tau_{yy}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ \frac{\partial}{\partial\tau_{zz}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ \frac{\partial}{\partial\tau_{xy}}\mathcal{I}_3(\boldsymbol{\tau})1 \\ \frac{\partial}{\partial\tau_{xz}}\mathcal{I}_3(\boldsymbol{\tau})1 \\ \frac{\partial}{\partial\tau_{yz}}\mathcal{I}_3(\boldsymbol{\tau})1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial}{\partial\tau_{xx}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ \frac{\partial}{\partial\tau_{yy}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ \frac{\partial}{\partial\tau_{zz}}\mathcal{I}_3(\boldsymbol{\tau})\frac{2}{3} \\ 2\tau_{xx}\tau_{xy} + 2\tau_{yy}\tau_{xy} + 2\tau_{xz}\tau_{yz} \\ 2\tau_{xx}\tau_{xz} + 2\tau_{zz}\tau_{xz} + 2\tau_{xy}\tau_{yz} \\ 2\tau_{yy}\tau_{yz} + 2\tau_{zz}\tau_{yz} + 2\tau_{xy}\tau{xz} \end{pmatrix} \tag{R.9}$$

$$= \begin{pmatrix} (\tau_{xx}^2 + \tau_{xy}^2 + \tau_{xz}^2)\frac{2}{3} \\ (\tau_{xy}^2 + \tau_{yy}^2 + \tau_{yz}^2)\frac{2}{3} \\ (\tau_{xz}^2 + \tau_{yz}^2 + \tau_{zz}^2)\frac{2}{3} \\ 2\tau_{xx}\tau_{xy} + 2\tau_{yy}\tau_{xy} + 2\tau_{xz}\tau_{yz} \\ 2\tau_{xx}\tau_{xz} + 2\tau_{zz}\tau_{xz} + 2\tau_{xy}\tau_{yz} \\ 2\tau_{yy}\tau_{yz} + 2\tau_{zz}\tau_{yz} + 2\tau_{xy}\tau{xz} \end{pmatrix} \tag{R.10}$$

We have $\boldsymbol{\tau} = \boldsymbol{\sigma} - \frac{1}{3}\mathcal{I}_1(\boldsymbol{\sigma})\mathbf{1}$ so

$$\tau_{xx} = \sigma_{xx} - \frac{1}{3}\mathcal{I}_1 \qquad\qquad \Rightarrow \tau_{xx}^2 = (\sigma_{xx} - \frac{1}{3}\mathcal{I}_1)^2 = \sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2 \qquad (\text{R.11})$$

$$\tau_{yy} = \sigma_{yy} - \frac{1}{3}\mathcal{I}_1 \qquad\qquad \Rightarrow \tau_{yy}^2 = (\sigma_{yy} - \frac{1}{3}\mathcal{I}_1)^2 = \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2 \qquad (\text{R.12})$$

$$\tau_{zz} = \sigma_{zz} - \frac{1}{3}\mathcal{I}_1 \qquad\qquad \Rightarrow \tau_{zz}^2 = (\sigma_{zz} - \frac{1}{3}\mathcal{I}_1)^2 = \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2 \qquad (\text{R.13})$$

Finally

$$\frac{\partial \mathcal{I}_3(\boldsymbol{\tau})}{\partial \vec{\sigma}} = \begin{pmatrix} (\sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2 + \sigma_{xy}^2 + \sigma_{xz}^2)\frac{2}{3} \\ (\sigma_{xy}^2 + \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2 + \sigma_{yz}^2)\frac{2}{3} \\ (\sigma_{xz}^2 + \sigma_{yz}^2 + \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1 + \frac{1}{9}\mathcal{I}_1^2)\frac{2}{3} \\ 2(\sigma_{xx} - \frac{1}{3}\mathcal{I}_1)\sigma_{xy} + 2(\sigma_{yy} - \frac{1}{3}\mathcal{I}_1)\sigma_{xy} + 2\sigma_{xz}\sigma_{yz} \\ 2(\sigma_{xx} - \frac{1}{3}\mathcal{I}_1)\sigma_{xz} + 2(\sigma_{zz} - \frac{1}{3}\mathcal{I}_1)\sigma_{xz} + 2\sigma_{xy}\sigma_{yz} \\ 2(\sigma_{yy} - \frac{1}{3}\mathcal{I}_1)\sigma_{yz} + 2(\sigma_{zz} - \frac{1}{3}\mathcal{I}_1)\sigma_{yz} + 2\sigma_{xy}\sigma xz \end{pmatrix} \qquad (\text{R.14})$$

$$= \begin{pmatrix} (\sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1 + \sigma_{xy}^2 + \sigma_{xz}^2)\frac{2}{3} \\ (\sigma_{xy}^2 + \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1 + \sigma_{yz}^2)\frac{2}{3} \\ (\sigma_{xz}^2 + \sigma_{yz}^2 + \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1)\frac{2}{3} \\ 2\sigma_{xx}\sigma_{xy} + 2\sigma_{yy}\sigma_{xy} + 2\sigma_{xz}\sigma_{yz} \\ 2\sigma_{xx}\sigma_{xz} + 2\sigma_{zz}\sigma_{xz} + 2\sigma_{xy}\sigma_{yz} \\ 2\sigma_{yy}\sigma_{yz} + 2\sigma_{zz}\sigma_{yz} + 2\sigma_{xy}\sigma xz \end{pmatrix} + \begin{pmatrix} \frac{1}{9}\mathcal{I}_1^2\frac{2}{3} \\ \frac{1}{9}\mathcal{I}_1^2\frac{2}{3} \\ \frac{1}{9}\mathcal{I}_1^2\frac{2}{3} \\ -\frac{4}{3}\mathcal{I}_1\sigma_{xy} \\ -\frac{4}{3}\mathcal{I}_1\sigma_{xz} \\ -\frac{4}{3}\mathcal{I}_1\sigma_{yz} \end{pmatrix}$$

$$= \begin{pmatrix} (\sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1 + \sigma_{xy}^2 + \sigma_{xz}^2)\frac{2}{3} \\ (\sigma_{xy}^2 + \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1 + \sigma_{yz}^2)\frac{2}{3} \\ (\sigma_{xz}^2 + \sigma_{yz}^2 + \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1)\frac{2}{3} \\ 2(\sigma_{xx} + \sigma_{yy})\sigma_{xy} + 2\sigma_{xz}\sigma_{yz} \\ 2(\sigma_{xx} + \sigma_{zz})\sigma_{xz} + 2\sigma_{xy}\sigma_{yz} \\ 2(\sigma_{yy} + \sigma_{zz})\sigma_{yz} + 2\sigma_{xy}\sigma xz \end{pmatrix} + \frac{2}{3}\mathcal{I}_1 \begin{pmatrix} \frac{1}{9}\mathcal{I}_1 \\ \frac{1}{9}\mathcal{I}_1 \\ \frac{1}{9}\mathcal{I}_1 \\ -2\sigma_{xy} \\ -2\sigma_{xz} \\ -2\sigma_{yz} \end{pmatrix} \qquad (\text{R.15})$$

$$= \begin{pmatrix} (\sigma_{xx}^2 - \frac{2}{3}\sigma_{xx}\mathcal{I}_1 + \sigma_{xy}^2 + \sigma_{xz}^2)\frac{2}{3} \\ (\sigma_{xy}^2 + \sigma_{yy}^2 - \frac{2}{3}\sigma_{yy}\mathcal{I}_1 + \sigma_{yz}^2)\frac{2}{3} \\ (\sigma_{xz}^2 + \sigma_{yz}^2 + \sigma_{zz}^2 - \frac{2}{3}\sigma_{zz}\mathcal{I}_1)\frac{2}{3} \\ 2(\mathcal{I}_1 - \sigma_{zz})\sigma_{xy} + 2\sigma_{xz}\sigma_{yz} \\ 2(\mathcal{I}_1 - \sigma_{yy})\sigma_{xz} + 2\sigma_{xy}\sigma_{yz} \\ 2(\mathcal{I}_1 - \sigma_{xx})\sigma_{yz} + 2\sigma_{xy}\sigma xz \end{pmatrix} + \frac{2}{3}\mathcal{I}_1 \begin{pmatrix} \frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{yy} & \frac{1}{3}\sigma_{zz} & & & \\ \frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{yy} & \frac{1}{3}\sigma_{zz} & & & \\ \frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{yy} & \frac{1}{3}\sigma_{zz} & & & \\ & & & -2 & & \\ & & & & -2 & \\ & & & & & -2 \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} \qquad (\text{R.16})$$

FINISH!!!!! not complicated but no rush. Also try to see whether it matches table I of Zienkiewicz

and Cormeau [1533] - note the different ordering of terms in vector

$$
\begin{pmatrix}
\frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{zz} & \frac{1}{3}\sigma_{yy} & -\frac{2}{3}\sigma_{yz} & \frac{1}{3}\sigma_{xz} & \frac{1}{3}\sigma_{xy} \\
\frac{1}{3}\sigma_{zz} & \frac{1}{3}\sigma_{yy} & \frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{yz} & -\frac{2}{3}\sigma_{xz} & \frac{1}{3}\sigma_{xy} \\
\frac{1}{3}\sigma_{yy} & \frac{1}{3}\sigma_{xx} & \frac{1}{3}\sigma_{zz} & \frac{1}{3}\sigma_{yz} & \frac{1}{3}\sigma_{xz} & -\frac{2}{3}\sigma_{xy} \\
-\frac{2}{3}\sigma_{yz} & \frac{1}{3}\sigma_{yz} & \frac{1}{3}\sigma_{yz} & -\sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\
\frac{1}{3}\sigma_{xz} & -\frac{2}{3}\sigma_{xz} & \frac{1}{3}\sigma_{xz} & \sigma_{xy} & -\sigma_{yy} & \sigma_{yz} \\
\frac{1}{3}\sigma_{xy} & \frac{1}{3}\sigma_{xy} & -\frac{2}{3}\sigma_{xy} & \sigma_{xz} & \sigma_{yz} & -\sigma_{zz}
\end{pmatrix}
\cdot
\begin{pmatrix}
\sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{xz} \\ \sigma_{xy}
\end{pmatrix}
\tag{R.17}
$$

$$
= \frac{1}{3}
\begin{pmatrix}
\sigma_{xx} & \sigma_{zz} & \sigma_{yy} & -2\sigma_{yz} & \sigma_{xz} & \sigma_{xy} \\
\sigma_{zz} & \sigma_{yy} & \sigma_{xx} & \sigma_{yz} & -2\sigma_{xz} & \sigma_{xy} \\
\sigma_{yy} & \sigma_{xx} & \sigma_{zz} & \sigma_{yz} & \sigma_{xz} & -2\sigma_{xy} \\
-2\sigma_{yz} & \sigma_{yz} & \sigma_{yz} & -3\sigma_{xx} & 3\sigma_{xy} & 3\sigma_{xz} \\
\sigma_{xz} & -2\sigma_{xz} & \sigma_{xz} & 3\sigma_{xy} & -3\sigma_{yy} & 3\sigma_{yz} \\
\sigma_{xy} & \sigma_{xy} & -2\sigma_{xy} & 3\sigma_{xz} & 3\sigma_{yz} & -3\sigma_{zz}
\end{pmatrix}
\cdot
\begin{pmatrix}
\sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{xz} \\ \sigma_{xy}
\end{pmatrix}
\tag{R.18}
$$

$$
=
\begin{pmatrix}
\sigma_{xx}^2 + 2\sigma_{yy}\sigma_{zz} - \sigma_{xy}\sigma_{yz} + \sigma_{xz}^2 \\
\sigma_{yy}^2 + 2\sigma_{xx}\sigma_{zz} + 2\sigma_{xy}\sigma_{yz} - 2\sigma_{xz}^2 \\
\sigma_{zz}^2 + 2\sigma_{xx}\sigma_{yy} - \sigma_{xy}\sigma_{yz} + \sigma_{xz}^2 \\
(-2\sigma_{xx} + \sigma_{yy} + \sigma_{zz})\sigma_{yz} - 3\sigma_{xx}\sigma_{xy} + 3\sigma_{xy}\sigma_{xz} + 3\sigma_{xz}\sigma_{yz} \\
(\sigma_{xx} - 2\sigma_{yy} + \sigma_{zz})\sigma_{xz} + 3\sigma_{xy}^2 - 3\sigma_{yy}\sigma_{xz} + 3\sigma_{yz}^2 \\
(\sigma_{xx} + \sigma_{yy} - 2\sigma_{zz})\sigma_{xy} \\
+3\sigma_{xy}\sigma_{xz} + 3\sigma_{xz}\sigma_{xz} - 3\sigma_{xz}\sigma_{yz}
\end{pmatrix}
\tag{R.19}
$$

Appendix S

Using gmsh

In the left menu *Modules>Geometry>Elementary entities>Add* choose *Rectangle* and input the coordinates of the lower left corner and its size. Then click on *Tools>Options* and *Tools>Statistics*. Your screen should look like this:
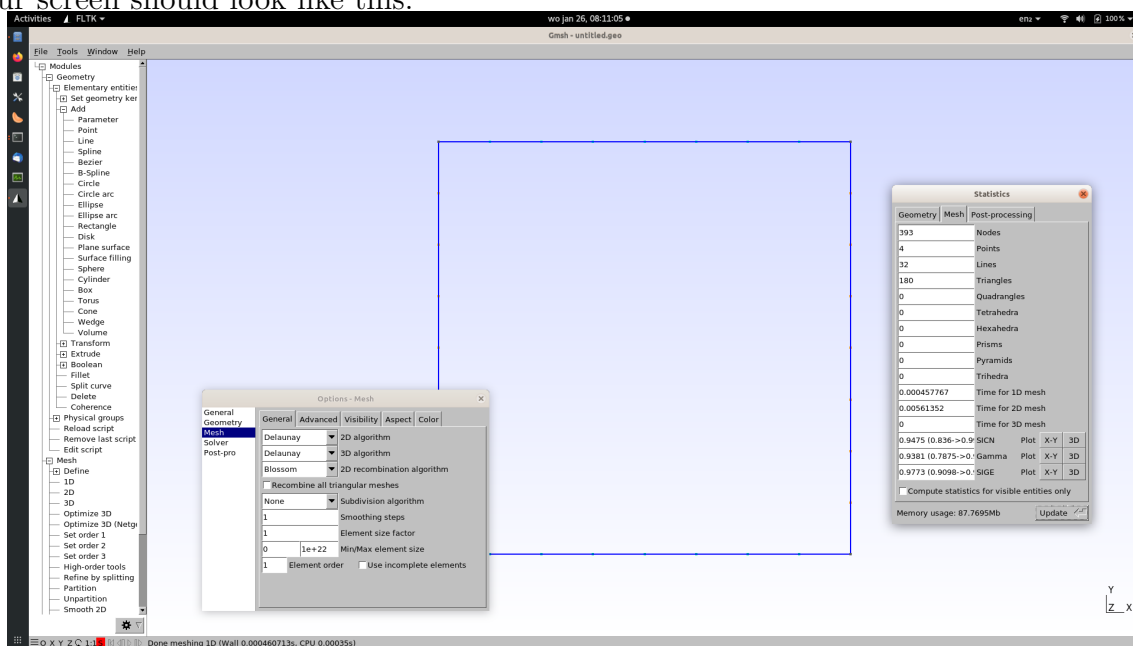
In the left menu click on *Mesh>2D* to generate an unstructured mesh. You should get this:
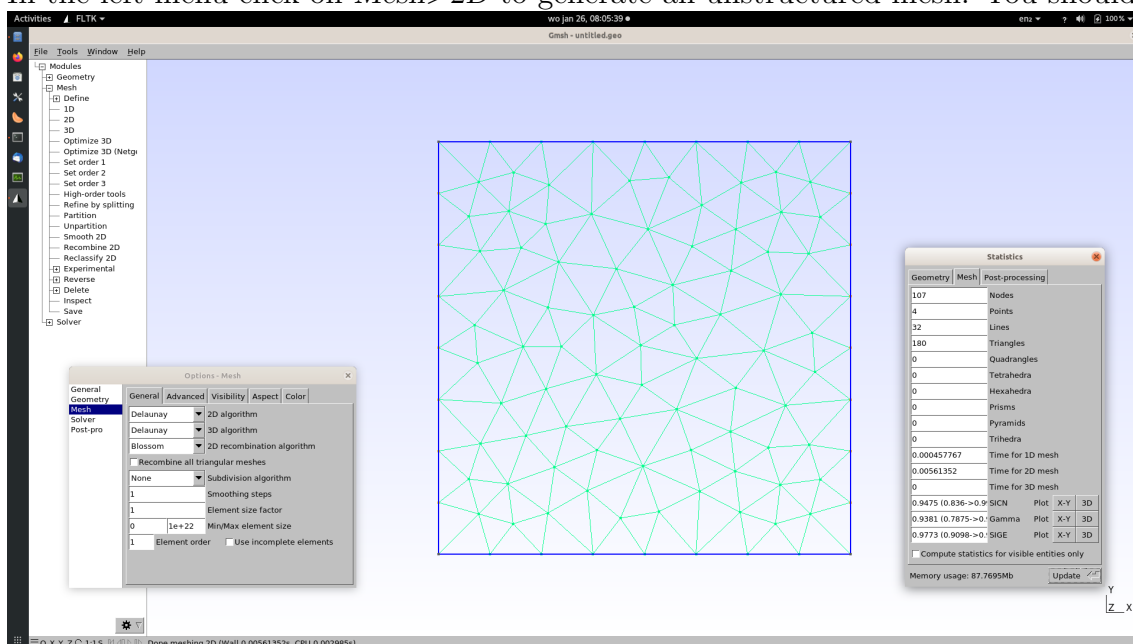
Make sure you can visualise the nodes by setting:
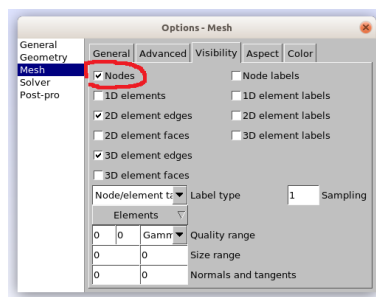
If you wish to use quadrilaterals
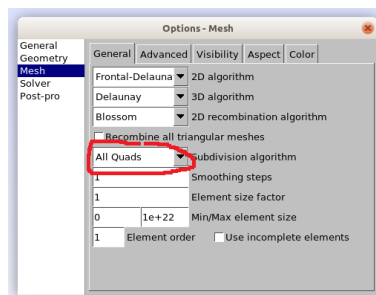
If you wish to generate second order elements click on *Mesh>Set order 2*:
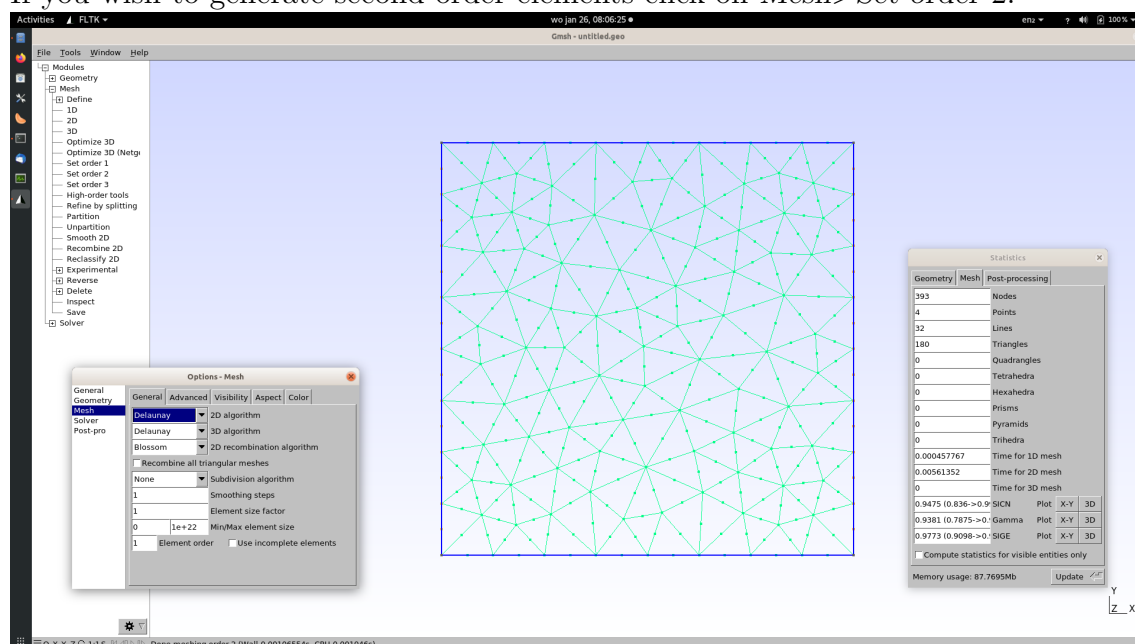
You can refine the mesh by clicking on *Refine by splitting*

Export the mesh: *File>Export*. Choose .mesh format.

In *Options - General - General tab* click on *Use dark interface* and in *Options - Mesh - Visibility tab* click on *Node labels*. Your screen now looks like this:
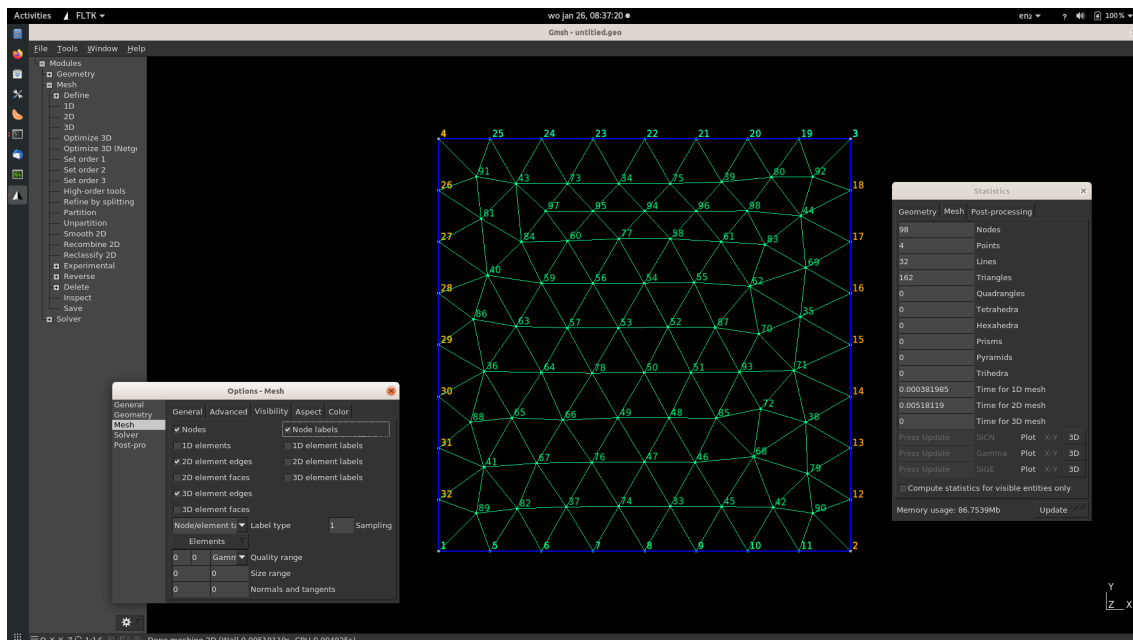
Also in Options - Mesh - General choose Delaunay instead of Frontal Delaunay to get this
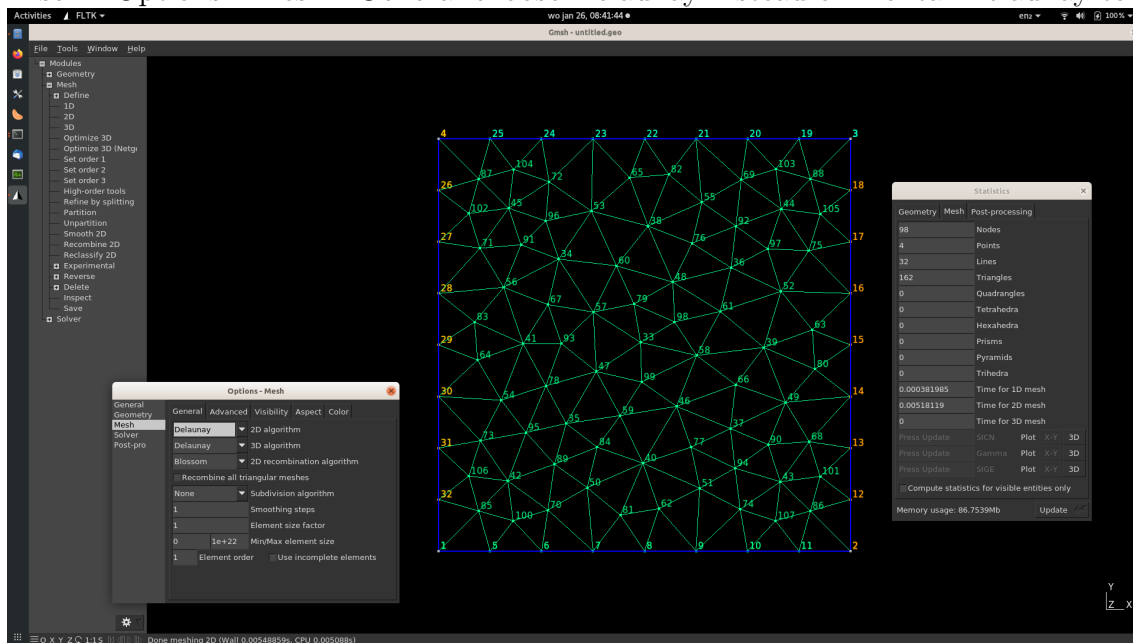
In order to re-generate a mesh, you mish wish to click on 1D and then on 2D again.

Appendix T

Directional derivative, total and material derivative

The following is a brief review of some basics of differential calculus which underlie many derivations in continuum mechanics.

T.0.1 Directional derivative

Let's start with a 2-dimensional example of a function f of the two variables x and y, hence $f(x, y)$.

Consider an arbitrary point (x_0, y_0) in the domain of f. We want to determine the rate of change of f in any direction in (x_0, y_0).

Let (x_1, y_1) be another point and define the unit vector points from (x_0, y_0) in the direction of (x_1, y_1) as:

$$\vec{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \frac{1}{d} \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix} \qquad \text{with} \qquad d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

The line segment connecting the two points can be parameterised as:

$$\begin{aligned} x &= x_0 + sn_1 \\ y &= y_0 + sn_2 \qquad , \quad s \in [0, d] \end{aligned} \tag{T.1}$$

Note that s, the arclength parameter, has the same dimension as the coordinates. On this line the function f is described by the 1-D function $f(s) = f(x(s), y(s))$, $s \in [0; d]$. The rate of change of $f(x, y)$ in the direction of $\vec{n}$ at some point $(x(s), y(s))$ on the line is then $\partial f(s) / \partial s$. This derivative can be related to the original coordinates as follows using the change rule of partial differentiation:

$$\frac{df(s)}{ds} = \frac{f(x(s), y(s))}{ds} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial s} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial s}$$

Using Eq. (T.1) this gives:

$$\frac{df(s)}{ds} = \frac{\partial f}{\partial x} n_1 + \frac{\partial f}{\partial y} n_2 = \vec{\nabla} f \cdot \vec{n} \tag{T.2}$$

This is the so-called directional derivative which can be computed in any point (x, y) and direction $\vec{n}$ as long as the two basic partial derivatives $\partial f / \partial x$ and $\partial f / \partial y$, which give the rate of change in the positive direction of the two axes, respectively, exist in (x, y).

Importantly, $df(s)/ds$ can be directly compared to $\partial f / \partial x$ and $\partial f / \partial y$ because all derivatives have the same physical dimension in any application by virtue of the parameterisation (T.1). A change of parameterisation parameter affects the l.h.s. of (T.2) but not the r.h.s because the latter depends

on the x and y coordinates and the unit vector n (which by definition (T.1) is dimensionless). A change of parameterisation variable from arc-length s to time $t = s/v$ will change the l.h.s. into

$$\frac{df}{ds} = \frac{df}{dt}\frac{dt}{ds} = \frac{1}{v}\frac{df}{dt}$$

Substituting this result in (T.2) and rewriting gives another type of directional derivative:

$$\frac{df(t)}{dt} = \vec{\nabla} \cdot \vec{v}$$

where $u\vec{n} = v\vec{n}$ can be interpreted as the local velocity vector, but only if this would be useful in the context of what f physically represents. Such re-parameterisation is useful in case one explicitly wants to determine the rate-of-change of f with respect to a parameter different from the arc-length s. The material derivative of continuum mechanics is an example of such a scaled directional derivative (see below).

T.0.2   Total differential

We can make the result (T.2) independent of parameterization as follows. By differentiating (T.1) to s, we find

$$dx(s) = n_1 ds \qquad \text{and} \qquad dy(s) = n_2 ds \qquad (T.3)$$

If we align the differential vector $d\vec{r} = (dx, dy)^T$ with the line segment we can write $d\vec{r} = (dx(s), dy(s))^T$. By using (T.3) we get $d\vec{r}(s) = \vec{n}ds$ and $|d\vec{r}| = ds = \sqrt{dx^2 + dy^2}$ because $\vec{n}$ is of unit length. These relations between $d\vec{r}$, dx, dy, and ds, all with the same physical dimension, are general used. Next, rewriting (T.2) as $df(s) = (\vec{\nabla} \cdot \vec{n})ds = \vec{\nabla} \cdot (\vec{n}ds)$ gives:

$$df(s) = \vec{\nabla}f \cdot d\vec{r}(s)$$

Because the line segment $(x_0, y_0) \rightarrow (x_1, y_1)$ is arbitrarily chosen we can as well write

$$df = \vec{\nabla}f \cdot d\vec{r} = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy \qquad (T.4)$$

Equation (T.4) is called the **total differential** of $f(x, y)$ which holds in each point (x, y) where the partial derivatives are calculated.

   This leads to the following interpretation: Given a function $f(x, y)$ then in any point (x, y) in which the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ exist we can compute the change df in f that occurs when going from $(x, y) \rightarrow (x + dx, y + dy)$ as (T.4), where $df = f(x + dx, y + dy) - f(x, y)$. This holds for every choice, including 0 or negative, of the differential steps dx and dy.

   Generalisation to N-dimensional space: For any multi-parameter function $f(x_1, ...x_N)$ equation (T.4) generalizes to the total differential

$$df = \frac{\partial f}{\partial x_1}dx_1 + ...\frac{\partial f}{\partial x_N}dx_N = \vec{\nabla}f^T \cdot d\vec{r}$$

Similarly, equations (T.1) and (T.2) can be generalized to functions of any number of parameters by parameterising the line connecting points $(x_1^0, ...x_N^0)$ and $(x_1, ...x_N)$:

$$x_1 = x_1^0 + sn_1$$
$$...$$
$$x_N = x_N^0 + sn_N$$

with $s \in [0; d]$ and $d = \sqrt{(x_1 - x_1^0)^2 + (x_N - x_N^0)^2}$.

Direction of maximal change: It follows from (T.2) or (T.4) that the change of a function is largest if $\vec{\nabla} f \cdot d\vec{r}$ is maximum which occurs in any chosen point when $\vec{\nabla} f$ is parallel to $d\vec{r}$. This implies that in every point the gradient vector $\vec{\nabla} f$ always points in the direction of maximum change of f and that $|\vec{\nabla} f| = |df(s)/ds|$ is that maximum change.

Calculating a normal vector: Suppose that $f(x_1, ..., x_N) = k$ is the level surface of function f for the constant k (for example the irregular and time-dependent temperature surface $T(t, x_1, x_2, x_3) = 20$ degrees in a room full of people). The equation $f(x_1, ..., x_N) = k$ implicitly defines the $(N-1)$-dimensional surface in N-dimensional space of all points for which $f = k$. We want to determine in any chosen point of this surface the vector $\vec{n}$ that is perpendicular to the surface. This is done as follows: Consider (T.4): $df = \vec{\nabla} f \cdot d\vec{r}$ and take $d\vec{r}$ to be a step from a point $(x_1, ...x_N)$ on $f = k$ along the level surface, i.e. $d\vec{r}$ lies in the level surface. In this case $df = 0$ because $f = k$ on the level surface. We find from (T.4) that $\vec{\nabla} f \cdot d\vec{r} = 0$, implying that $\vec{\nabla} f$ is perpendicular to $d\vec{r}$. Hence, $\vec{\nabla} f$ is a vector which is always normal to a level surface. The unit normal in any point $(x_1, ...x_N)$ on the level surface is then calculated as: $\vec{n} = \vec{\nabla} f / |\vec{\nabla} f|$ where $\vec{\nabla} f$ is the gradient in that point.

A corollary of this result is that if one calculates $\vec{\nabla} f$ in some point $(x_1^0, ...x_N^0)$ in N-space, then one also knows the local direction of the level surface $f(x_1, ...x_N) = f(x_1^0, ...x_N^0)$ that passes through $(x_1^0, ...x_N^0)$.

T.0.3 The material derivative

In continuum mechanics we distinguish the spatial coordinates x_1, x_2, x_3 and time t. Hence any function defined on this 4-parameter space is written as $f(t, x_1, x_2, x_3)$. The total differential (10) is then

$$df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \frac{\partial f}{\partial x_3} dx_3 \tag{T.5}$$

In principle, the time differential dt and spatial differentials dx_i can be arbitrarily chosen. For instance, taking $dt = 0$ that only the spatial changes in the function at fixed time are considered, while taking $dx_i = 0$ focuses on the temporal variation in a chosen fixed point. Generally, in continuum mechanics a special choice is made for the directional derivative, which involves the local direction of the flow. This direction is given at any point (x_1, x_2, x_3) and any time t by the velocity vector

$$\vec{v}(t, x_1, x_2, x_3) = \frac{d\vec{r}}{dt} \tag{T.6}$$

where $d\vec{r}$ is the spatial step taken by a flow particle from $(x_1, x_2, x_3) \to (x_1 + dx_1, x_2 + dx_2, x_3 + dx_3)$ during the time interval $t \to t + dt$. Hence, the flow direction $d\vec{r}$ in point (x_1, x_2, x_3) depends on the time t such that $d\vec{r}(t) = v(t, x_1, x_2, x_3)dt$.

Taking $\vec{v} = (v_1, v_2, v_3)^T$ equation (T.5) becomes

$$df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x_1} v_1 dt + \frac{\partial f}{\partial x_2} v_2 dt + \frac{\partial f}{\partial x_3} v_3 dt$$

Dividing by δt yields

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x_1} v_1 + \frac{\partial f}{\partial x_2} v_2 + \frac{\partial f}{\partial x_3} v_3 == \frac{\partial f}{\partial t} + \vec{v} \cdot \vec{\nabla} f$$

This equation is called the material derivative of f and describes the rate of change of f with time in the local direction of the flow.

FINISH

T.0.4 Material derivative of a volume integral

Let $F(\vec{r}, t)$ be some scalar function depending on spatial coordinates $\vec{r}$ and time t and $V(t)$ a volume that may also depend on t. Define the volume integral

$$I(t) = \int_{V(t)} F(\vec{r}, t) dV.$$

For example, if F is density, then $I(t)$ is the mass contained in the volume.

Assume a deforming medium with incremental displacement field $\vec{s}(\vec{r}, t)$. Consider the deformation that occurs between t and $t + \Delta t$ in which Δt is a very small time step such that we can write that a particle at position $\vec{r}$ at time t will be displaced to $\vec{r} + \Delta \vec{r}$ at $t + \Delta t$.

Then

$$\Delta \vec{r} = \vec{s}(\vec{r}, t + \Delta t) - \vec{s}(\vec{r}, t) = \frac{\vec{s}(\vec{r}, t + \Delta t) - \vec{s}(\vec{r}, t)}{\Delta t} \Delta t = \vec{v} \Delta t$$

where $\vec{v} = d\vec{s}/dt$ is the velocity vector at $(\vec{r}, t)$.

The volume $V(t)$ will deform to $V'(t + \Delta t)$. The material derivative of $I(t)$ is defined as:

$$\frac{DI}{Dt} = \frac{D}{Dt} \int_{V(t)} F(\vec{r}, t) dV = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\int_{V'(t+\Delta t)} F(\vec{r} + \Delta \vec{r}, t + \Delta t) dV' - \int_{V(t)} F(\vec{r}, t) dV \right] \quad \text{(T.7)}$$

For incremental Δt we can approximate

$$F(\vec{r} + \Delta \vec{r}, t + \Delta t) = F(\vec{r}, t) + \frac{\partial F}{\partial x_j} v_j \Delta t + \frac{\partial F}{\partial t} \Delta t = F(\vec{r}, t) + \frac{DF}{Dt} \Delta t$$

Further, from continuum mechanics we have for the volume change associated with the incremental displacement field $\vec{s}(\vec{r}, t)$:

$$\frac{dV' - dV}{dV'} = \vec{\nabla} \cdot \vec{s} = \vec{\nabla} \cdot (\vec{v} \Delta t)$$

or

$$dV' = \left(1 + \frac{\partial v_j}{\partial x_j} \Delta t \right) dV$$

Using these results

$$F(\vec{r} + \Delta \vec{r}, t + \Delta t) dV' = \left(F(\vec{r}, t) + \frac{DF}{Dt} \Delta t \right) \left(1 + \frac{\partial v_j}{\partial x_j} \Delta t \right) dV = F(\vec{r}, t) dV + \frac{DF}{Dt} \Delta t dV + F(\vec{r}, t) \frac{\partial v_j}{\partial x_j} \Delta t dV + \frac{DF}{Dt}$$

such that now the integration over V' can be replaced by an integration over V:

$$\int_{V'(t+\delta t)} F(\vec{r} + \Delta \vec{r}, t + \Delta t) dV' \simeq \int_{V(t)} F(\vec{r}, t) dV + \int_{V(t)} \left[\left(\frac{DF}{Dt} + \frac{\partial v_j}{\partial x_j} \right) \Delta t + \frac{DF}{Dt} \frac{\partial v_j}{\partial x_j} \Delta t^2 \right] dV$$

Substituting this result in the above definition of the material derivative

$$\frac{DI}{Dt} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\int_{V(t)} \left(\frac{DF}{Dt} + \frac{\partial v_j}{\partial x_j} \right) \Delta t + \frac{DF}{Dt} \frac{\partial v_j}{\partial x_j} \Delta t^2 \quad dV \right]$$

This leads to material derivative of a volume integral:

$$\frac{D}{Dt} \int_{V(t)} F fV = \int_{V(t)} \frac{DF}{Dt} + F \frac{\partial v_j}{\partial x_j} \quad dV$$

Bibliography

[1] B.T. Aagaard, M.G. Knepley, and C.A. Williams. "A domain decomposition approach to implementing fault slip in finite-element models of quasi-static and dynamic crustal deformation". In: *J. Geophys. Res.: Solid Earth* 118 (2013), pp. 3059–3079. DOI: 10.1002/jgrb.50217.

[2] F.J Adewale, A.P. Lucky, A.P. Oluwabunmi, and E.F. Boluwaji. "Selecting the most appropriate model for rheological characterization of synthetic based drilling mud". In: *International Journal of Applied Engineering Research* 12.18 (2017), pp. 7614–7649. DOI: xxxx.

[3] Deniz Tolga Akcabay, David R Dowling, and William W Schultz. "Clothes washing simulations". In: *Computers & Fluids* 100 (2014), pp. 79–94. DOI: 10.1016/j.compfluid.2014.05.005.

[4] Francis Albarède and Rob D van der Hilst. "Zoned mantle convection". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360.1800 (2002), pp. 2569–2592. DOI: 10.1098/rsta.2002.1081.

[5] M. Albers. "A local mesh refinement multigrid method for 3D convection problems with strongly variable viscosity". In: *J. Comp. Phys.* 160 (2000), pp. 126–150.

[6] Leandro R Alejano and Antonio Bobet. "Drucker–prager criterion". In: *The ISRM Suggested Methods for Rock Characterization, Testing and Monitoring: 2007-2014.* 2012, pp. 247–252. DOI: 10.1007/s00603-012-0278-2.

[7] L. Alisic, M. Gurnis, G. Stadler, C. Burstedde, and O. Ghattas. "Multi-scale dynamics and rheology of mantle flow with plates". In: *J. Geophys. Res.: Solid Earth* 117 (2012). DOI: 10.1029/2012JB009234.

[8] Richard B Alley. "Flow-law hypotheses for ice-sheet modeling". In: *Journal of Glaciology* 38.129 (1992), pp. 245–256. DOI: 10.3189/S0022143000003658.

[9] V. Allken, R. Huismans, and C. Thieulot. "Factors controlling the mode of rift interaction in brittle-ductile coupled systems: a 3D numerical study". In: *Geochem. Geophys. Geosyst.* 13.5 (2012), Q05010. DOI: 10.1029/2012GC004077.

[10] V. Allken, R. Huismans, and C. Thieulot. "Three dimensional numerical modelling of upper crustal extensional systems". In: *J. Geophys. Res.: Solid Earth* 116 (2011), B10409. DOI: 10.1029/2011JB008319.

[11] V. Allken, R.S. Huismans, H. Fossen, and C. Thieulot. "3D numerical modelling of graben interaction and linkage: a case study of the Canyonlands grabens, Utah". In: *Basin Research* 25 (2013), pp. 1–14. DOI: 10.1111/bre.12010.

[12] P.R. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary. "Performance and Scalability of the Block Low-Rank Multifrontal Factorization on Multicore Architectures". In: *ACM Transactions on Mathematical Software* 45 (1 2019), 2:1–2:26. DOI: 10.1145/3242094.

[13] P.R. Amestoy and I.S. Duff. "Vectorization of a multiprocessor multifrontal code". In: *International Journal of Supercomputer Applications* 3 (1989), pp. 41–59. DOI: `10.1177/109434208900300303`.

[14] P.R. Amestoy, I.S. Duff, J.Koster, and J.-Y. L'Excellent. "A fully asynchronous multifrontal solver using distributed dynamic scheduling". In: *SIAM Journal of Matrix Analysis and Applications* 23.1 (2001), pp. 15–41.

[15] P.R. Amestoy, I.S. Duff, and J.-Y. L'Excellent. "Multifrontal parallel distributed symmetric and unsymmetric solvers". In: *Computer Methods in Applied Mechanics and Engineering* 184 (2000), pp. 501–520.

[16] P.R. Amestoy, A. Guermouche, J.-Y. L'Excellent, and S. Pralet. "Hybrid scheduling for the parallel solution of linear systems". In: *Parallel Computing* 32.2 (2006), pp. 136–156. DOI: `10.1016/j.parco.2005.07.004`.

[17] Christophe Ancey and Steve Cochard. "The dam-break problem for Herschel–Bulkley viscoplastic fluids down steep flumes". In: *Journal of Non-Newtonian Fluid Mechanics* 158.1-3 (2009), pp. 18–35. DOI: `10.1016/j.jnnfm.2008.08.008`.

[18] Edward Anders and Nicolas Grevesse. "Abundances of the elements: Meteoritic and solar". In: *Geochimica et Cosmochimica acta* 53.1 (1989), pp. 197–214. DOI: `10.1016/0016-7037(89)90286-X`.

[19] B.D. Anderson. "AUTOMATED ALL-QUADRILATERAL MESH ADAPTATION THROUGH REFINEMENT AND COARSENING". PhD thesis. Department of Civil and Environmental Engineering, Brigham Young University, 2009.

[20] Bret D Anderson, Steven E Benzley, and Steven J Owen. "Automatic all quadrilateral mesh adaption through refinement and coarsening". In: *Proceedings of the 18th international meshing roundtable*. Springer, 2009, pp. 557–574. DOI: `10.1007/978-3-642-04319-2_32`.

[21] C.A. Anderson and R.J. Bridwell. "A finite element method for studying the transient nonlinear thermal creep of geological structures". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 4 (1980), pp. 255–276.

[22] J.D. Anderson. *Computational Fluid Dynamics*. McGraw-Hill, 1995.

[23] Thomas Andolfsson. "Analyses of thermal conductivity from mineral composition and analyses by use of Thermal Conductivity Scanner: a study of thermal properties of Scanian rock types". In: *MSc thesis - Dissertations in Geology at Lund University* (2013).

[24] Miguel Andrés-Martínez, Jason P Morgan, Marta Pérez-Gussinyé, and Lars Rüpke. "A new free-surface stabilization algorithm for geodynamical modelling: Theory and numerical tests". In: *Phys. Earth. Planet. Inter.* 246 (2015), pp. 41–51. DOI: `10.1016/j.pepi.2015.07.003`.

[25] Miguel Andrés-Martínez, Marta Pérez-Gussinyé, John Armitage, and Jason P Morgan. "Thermomechanical Implications of Sediment Transport for the Architecture and Evolution of Continental Rifts and Margins". In: *Tectonics* 38.2 (2019), pp. 641–665. DOI: `10.1029/2018TC005346`.

[26] D.N. Arnold, F. Brezzi, and M. Fortin. "A stable finite element for the Stokes equation". In: *Calcolo* XXI.IV (1984), pp. 337–344. DOI: `10.1007/BF02576171`.

[27] Douglas Arnold, Richard Falk, and Ragnar Winther. "Preconditioning in H(div) and applications". In: *Mathematics of computation* 66.219 (1997), pp. 957–984. DOI: `xxx`.

[28] Douglas N Arnold. "On nonconforming linear-constant elements for some variants of the Stokes equations". In: *Istit. Lombardo Accad. Sci. Lett. Rend. A* 127.1 (1993). DOI: `xxxx`.

[29] Douglas N Arnold, Daniele Boffi, and Richard S Falk. "Quadrilateral H(div) finite elements". In: *SIAM Journal on Numerical Analysis* 42.6 (2005), pp. 2429–2451. DOI: 10.1137/S0036142903431924.

[30] P.A. Arrial, N. Flyer, G.B. Wright, and L.H. Kellogg. "On the sensitivity of 3-D thermal convection codes to numerical discretization: a model intercomparison". In: *Geosci. Model. Dev.* 7 (2014), pp. 2065–2076. DOI: 10.5194/gmd-7-2065-2014.

[31] M. Arroyo, C. Couder-Castaneda, A. Trujillo-Alcantara, I.E. Herrera-Diaz, and N. Vera-Chavez. "A Performance Study of a Dual Xeon-Phi Cluster for the Forward Modelling of Gravitational Fields". In: *Scientific Programming* (2015). DOI: 10.1155/2015/316012.

[32] Alireza Asgari and LN Moresi. "Multiscale particle-in-cell method: from fluid to solid mechanics". In: *Advanced Methods for Practical Applications in Fluid Mechanics. InTech, Croatia* (2012), pp. 185–208.

[33] M.F. Asgharzadeh, R.R.B. Von Frese, H.R. Kim, T.E. Leftwich, and J.W. Kim. "Spherical prism gravity effects by Gauss-Legendre quadrature integration". In: *Geophy. J. Int.* 169.1 (2007), pp. 1–11. DOI: 10.1111/j.1365-246X.2007.03214.x.

[34] F. Auricchio, L. Beirão da Veiga, F. Brezzi, and C. Lovadina. "Mixed Finite Element Methods". In: *Encyclopedia of Computational Mechanics Second Edition.* 2017, pp. 1–53. ISBN: 9781119176817. DOI: 10.1002/9781119176817.ecm2004.

[35] C Auth and H Harder. "Multigrid solution of convection problems with strongly variable viscosity". In: *Geophy. J. Int.* 137.3 (1999), pp. 793–804. DOI: 10.1046/j.1365-246x.1999.00833.x.

[36] EH Ayachour. "A fast implementation for GMRES method". In: *Journal of Computational and Applied Mathematics* 159.2 (2003), pp. 269–283.

[37] S. Azuma and I. Katayama. "Evolution of the rheological structure of Mars". In: *Earth, Planets and Space* 69.1 (2017), pp. 1–13. DOI: 10.1186/s40623-016-0593-z.

[38] A Yu Babeyko, Stephan V Sobolev, RB Trumbull, Onno Oncken, and LL Lavier. "Numerical models of crustal scale convection and partial melting beneath the Altiplano–Puna plateau". In: *Earth Planet. Sci. Lett.* 199.3-4 (2002), pp. 373–388. DOI: 10.1016/S0012-821X(02)00597-6.

[39] Andrey Y Babeyko, Stephan V Sobolev, Tim Vietor, Onno Oncken, and Robert B Trumbull. "Numerical study of weakening processes in the central Andean back-arc". In: *The Andes.* 2006, pp. 495–512. DOI: 10.1007/978-3-540-48684-8_24.

[40] Ivo Babuška. "Error-bounds for finite element method". In: *Numerische Mathematik* 16.4 (1971), pp. 322–333. DOI: 10.1007/BF02165003.

[41] Corinna Bahriawati and Carsten Carstensen. "Three MATLAB implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control". In: *Computational methods in applied mathematics* 5.4 (2005), pp. 333–361. DOI: 10.2478/cmam-2005-0016.

[42] W. Bai. "The quadrilateral Mini finite element for the Stokes problem". In: *Computer Methods in Applied Mechanics and Engineering* 143 (1997), pp. 41–47. DOI: 10.1016/S0045-7825(96)01146-2.

[43] R.R. Bakker, M. Frehner, and M. Lupi. "How temperature-dependent elasticity alters host rock/magmatic reservoir models: A case study on the effects of ice-cap unloading on shallow volcanic systems". In: *Earth Planet. Sci. Lett.* 456 (2016), pp. 16–25. DOI: 10.1016/j.epsl.2016.09.039.

[44] N.J. Balmforth, Y. Forterre, and O. Pouliquen. "The viscoplastic Stokes layer". In: *Journal of Non-Newtonian Fluid Mechanics* 158 (2009), pp. 46–53. DOI: `10.1016/j.jnnfm.2008.07.008`.

[45] N.J. Balmforth and A.C. Rust. "Weakly nonlinear viscoplastic convection". In: *Journal of Non-Newtonian Fluid Mechanics* 158 (2009), pp. 36–45. DOI: `10.1016/j.jnnfm.2008.07.012`.

[46] NJ Balmforth, RV Craster, P Perona, AC Rust, and R Sassi. "Viscoplastic dam breaks and the Bostwick consistometer". In: *Journal of non-newtonian fluid mechanics* 142.1-3 (2007), pp. 63–78. DOI: `10.1016/j.jnnfm.2006.06.005`.

[47] Wolfgang Bangerth, Imbunm Kim, Dongwoo Sheen, and Jaeryun Yim. "On Hanging Node Constraints for Nonconforming Finite Elements using the Douglas–Santos–Sheen–Ye Element as an Example". In: *SIAM Journal on Numerical Analysis* 55.4 (2017), pp. 1719–1739. DOI: `10.1137/16M1071432`.

[48] Wolfgang Bangerth et al. ASPECT*: Advanced Solver for Problems in Earth's ConvecTion, User Manual.* doi:10.6084/m9.figshare.4865333. 2022. DOI: `10.6084/m9.figshare.4865333`. URL: `https://doi.org/10.6084/m9.figshare.4865333`.

[49] H.A. Barnes. "The yield stress - everything flows?" In: *J. Non-Newtonian Fluid Mech.* 81 (1999), pp. 133–178.

[50] HA Barnes and K Walters. "The yield stress myth?" In: *Rheologica acta* 24.4 (1985), pp. 323–326.

[51] Terence D Barr and Gregory A Houseman. "Deformation fields around a fault embedded in a non-linear ductile medium". In: *Geophy. J. Int.* 125.2 (1996), pp. 473–490. DOI: `10.1111/j.1365-246X.1996.tb00012.x`.

[52] R. Barrett et al. *Templates for the solution of linear systems: building blocks for iterative methods.* SIAM, 1994.

[53] TL Barry et al. "Whole-mantle convection with tectonic plates preserves long-term global patterns of upper mantle geochemistry". In: *Scientific Reports* 7.1 (2017), pp. 1–9. DOI: `10.1038/s41598-017-01816-y`.

[54] T. Barth, P. Bochev, M. Gunzburger, and J. Shadid. "A taxonomy of consistently stabilised finite element methods for the Stokes problem". In: *SIAM J. Sci. Comput.* 25.5 (2004), pp. 1585–1607. DOI: `10.1137/S1064827502407718`.

[55] PJ Barton. "The relationship between seismic velocity and density in the continental crust - a useful constraint?" In: *Geophy. J. Int.* 87.1 (1986), pp. 195–208.

[56] G.K. Batchelor. *An introduction to fluid dynamics.* Cambridge University Press, 1967.

[57] K.-J. Bathe. *Finite Element Procedures in Engineering Analysis.* Prentice-Hall, 1982.

[58] GE Batt and J Braun. "On the thermomechanical evolution of compressional orogens". In: *Geophy. J. Int.* 128.2 (1997), pp. 364–382. DOI: `10.1111/j.1365-246X.1997.tb01561.x`.

[59] L. Battaglia, M.A. Storti, and J. D'Elia. "An interface capturing finite element approach for free surface flows using unstructured grids". In: *Mecanica Computacional* XXVII (2008), pp. 33–48. DOI: `xxxxx`.

[60] Simon Bauer, Markus Huber, Siavash Ghelichkhan, Marcus Mohr, Ulrich Rüde, and B Wohlmuth. "Large-scale simulation of mantle convection based on a new matrix-free approach". In: *Journal of Computational Science* 31 (2019), pp. 60–76. DOI: `10.1016/j.jocs.2018.12.006`.

[61] Simon Bauer et al. "TerraNeo—Mantle Convection Beyond a Trillion Degrees of Freedom". In: *Software for Exascale Computing - SPPEXA 2016-2019*. Ed. by Hans-Joachim Bungartz, Severin Reiz, Benjamin Uekermann, Philipp Neumann, and Wolfgang E. Nagel. Springer International Publishing, 2020, pp. 569–610. DOI: 10.1007/978-3-030-47956-5_19.

[62] J.R. Baumgardner. "Three-Dimensional treatment of convective flow in the Earth's mantle". In: *Journal of Statistical Physics* 39.5/6 (1985), pp. 501–511. DOI: 10.1007/BF01008348.

[63] J.R. Baumgardner and P.O. Frederickson. "Isocahedral discretisation of the two-sphere". In: *SIAM J. Numer Anal.* 22.6 (1985), pp. 1107–1115. DOI: 10.1137/0722066.

[64] C. Beaumont, P. Fullsack, and J. Hamilton. "Erosional control of active compressional orogens". In: *Thrust Tectonics* 99 (1992), pp. 1–18. DOI: 10.1007/978-94-011-3066-0_1.

[65] C. Beaumont, P. Fullsack, and J. Hamilton. "Styles of crustal deformation in compressional orogens caused by subduction of the underlying lithosphere". In: *Tectonophysics* 232 (1994), pp. 119–132. DOI: 10.1016/0040-1951(94)90079-5.

[66] C. Beaumont, J.A. Munoz, J. Hamilton, and P. Fullsack. "Factors controlling the Alpine evolution of the central Pyrenees inferred from a comparison of observations and geodynamical models". In: *J. Geophys. Res.: Solid Earth* 105 (2000), pp. 8121–8145. DOI: 10.1029/1999JB900390.

[67] Christopher Beaumont and Anthony Lambert. "Crustal structure from surface load tilts, using a finite element model". In: *Geophy. J. Int.* 29.2 (1972), pp. 203–226.

[68] Christopher Beaumont and Garry Quinlan. "A geodynamic framework for interpreting crustal-scale seismic-reflectivity patterns in compressional orogens". In: *Geophy. J. Int.* 116.3 (1994), pp. 754–783. DOI: 10.1111/j.1365-246X.1994.tb03295.x.

[69] J.M. Becker and M. Bevis. "Love's problem". In: *Geophy. J. Int.* 156 (2004), pp. 171–178. DOI: 10.1111/j.1365-246X.2003.02150.x.

[70] T.W. Becker. "On the effect of temperature and strain-rate dependent viscosity on global mantle flow, net rotation, and plate-driving forces". In: *Geophy. J. Int.* 167 (2006), pp. 943–957.

[71] T.W. Becker and B.J.P. Kaus. *Numerical Geodynamics. v1.1.* Tech. rep. University Southern California, 2010.

[72] Thorsten W Becker and Richard J O'Connell. "Predicting plate velocities with mantle circulation models". In: *Geochem. Geophys. Geosyst.* 2.12 (2001).

[73] Mark D Behn, David L Goldsby, and Greg Hirth. "The role of grain size evolution in the rheology of ice: implications for reconciling laboratory creep data and the Glen flow law". In: *The Cryosphere* 15.9 (2021), pp. 4589–4605. DOI: 10.5194/tc-15-4589-2021.

[74] M. Behr. "On the Application of Slip Boundary Condition on Curved Boundaries". In: *Int. J. Num. Meth. Fluids* 45 (2004), pp. 43–51. DOI: 10.1002/fld.663.

[75] Léa Bello, Nicolas Coltice, Tobias Rolf, and Paul J Tackley. "On the predictability limit of convection models of the Earth's mantle". In: *Geochem. Geophys. Geosyst.* 15.6 (2014), pp. 2319–2328. DOI: 10.1002/2014GC005254.

[76] David J Benson. "An efficient, accurate, simple ALE method for nonlinear finite element programs". In: *Computer Methods in Applied Mechanics and Engineering* 72.3 (1989), pp. 305–350.

[77] M. Benzi, G.H. Golub, and J. Liesen. "Numerical solution of saddle point problems". In: *Acta Numerica* 14 (2005), pp. 1–137. DOI: 10.1017/S0962492904000212.

[78] M. Benzi and A.J. Wathen. "Some Preconditioning Techniques for Saddle Point Problems". In: *Model Order Reduction: Theory, Research Aspects and Applications*. Ed. by W.H.A. Schilders, H.A. van der Vorst, and Joost Rommes. Springer, 2008, pp. 195–211.

[79] D. Bercovici and G. Schubert. *Treatise on geophysics: Mantle dynamics. Vol. 7*. Elsevier, 2007.

[80] D. Bercovici and G. Schubert. *Treatise on geophysics: Mantle dynamics. Vol. 7 - Second Edition*. Elsevier, 2015.

[81] David Bercovici. "A simple model of plate generation from mantle flow". In: *Geophy. J. Int.* 114.3 (1993), pp. 635–650. DOI: 10.1111/j.1365-246X.1993.tb06993.x.

[82] David Bercovici. "A source-sink model of the generation of plate tectonics from non-Newtonian mantle flow". In: *J. Geophys. Res.: Solid Earth* 100.B2 (1995), pp. 2013–2030.

[83] David Bercovici. "Plate generation in a simple model of lithosphere-mantle flow with dynamic self-lubrication". In: *Earth Planet. Sci. Lett.* 144.1-2 (1996), pp. 41–51.

[84] M. Bercovier and M. Engelman. "A finite-element for the numerical solution of viscous incompressible flows". In: *J. Comp. Phys.* 30 (1979), pp. 181–201. DOI: 10.1016/0021-9991(79)90098-6.

[85] M. Bercovier and M. Engelman. "A finite-element method for incompressible Non-Newtonian flows". In: *J. Comp. Phys.* 36 (1980), pp. 313–326. DOI: 10.1016/0021-9991(80)90163-1.

[86] Luca Bergamaschi, Stefano Mantica, and Fausto Saleri. "Mixed finite element approximation of Darcy's law in porous media". In: *Report CRS4 AppMath-94-20, CRS4. Cagliari, Italy* 21 (1994).

[87] Christine Bernardi and Genevieve Raugel. "Analysis of Some Finite Elements for the Stokes Problem". In: *Mathematics of Computation* 44.169 (1985), pp. 71–79.

[88] Constance M Bertka and Yingwei Fei. "Density profile of an SNC model Martian interior and the moment-of-inertia factor of Mars". In: *Earth Planet. Sci. Lett.* 157.1-2 (1998), pp. 79–88. DOI: 10.1016/S0012-821X(98)00030-2.

[89] FH Bertrand, MR Gadbois, and PA Tanguy. "Tetrahedral elements for fluid flow". In: *Int. J. Num. Meth. Eng.* 33.6 (1992), pp. 1251–1267. DOI: 10.1002/nme.1620330610.

[90] B. Deglo de Besses, A. Magnin, and P. Jay. "Sphere drag in a viscoplastic fluid". In: *AIChE Journal* 50.10 (2004), pp. 2627–2629. DOI: 10.1002/aic.10252.

[91] Jack D Betteridge, Patrick E Farrell, and David A Ham. "Code generation for productive, portable, and scalable finite element simulation in firedrake". In: *Computing in Science & Engineering* 23.4 (2021), pp. 8–17. DOI: 10.1109/MCSE.2021.3085102.

[92] M.J. Beuchert and Y.Y. Podladchikov. "Viscoelastic mantle convection and lithospheric stresses". In: *Geophy. J. Int.* 183 (2010), pp. 35–63. DOI: 10.1111/j.1365-246X.2010.04708.x.

[93] Marcus J Beuchert, Yuri Y Podladchikov, Nina SC Simon, and Lars H Rüpke. "Modeling of craton stability using a viscoelastic rheology". In: *J. Geophys. Res.: Solid Earth* 115.B11 (2010). DOI: 10.1029/2009JB006482.

[94] Marco Bianchi and Daniele Pedretti. "An entrogram-based approach to describe spatial heterogeneity with applications to solute transport in porous media". In: *Water Resources Research* 54.7 (2018), pp. 4432–4448. DOI: 10.1029/2018WR022827.

[95] E.C. Bingham. *Fluidity and Plasticity*. McGraw-Hill, New York, 1922.

[96] Francis Birch. "Elasticity and constitution of the Earth's interior". In: *J. Geophys. Res.: Solid Earth* 57.2 (1952), pp. 227–286. DOI: `10.1029/JZ057i002p00227`.

[97] P. Bird and D.A. Yuen. "The use of the minimum-dissipation principle in tectonophysics". In: *Earth Planet. Sci. Lett.* 45 (1979), pp. 214–217. DOI: `10.1016/0012-821X(79)90122-5`.

[98] Peter Bird. "Finite element modeling of lithosphere deformation: the Zagros collision orogeny". In: *Tectonophysics* 50.2-3 (1978), pp. 307–336.

[99] Peter Bird. "Thin-plate and thin-shell finite-element programs for forward dynamic modeling of plate deformation and faulting". In: *Computers & Geosciences* 25.4 (1999), pp. 383–394.

[100] D. Bittner and H. Schmeling. "Numerical modelling of melting processes and induced diapirism in the lower crust". In: *Geophy. J. Int.* 123 (1995), pp. 59–70. DOI: `10.1111/j.1365-246X.1995.tb06661.x`.

[101] J Blackery and E Mitsoulis. "Creeping motion of a sphere in tubes filled with a Bingham plastic material". In: *Journal of non-newtonian fluid mechanics* 70.1-2 (1997), pp. 59–77.

[102] B. Blankenbach et al. "A benchmark comparison for mantle convection codes". In: *Geophy. J. Int.* 98 (1989), pp. 23–38. DOI: `10.1111/j.1365-246X.1989.tb05511.x`.

[103] A Blazquez, B Meyssignac, JM Lemoine, Etienne Berthier, A Ribes, and A Cazenave. "Exploring the uncertainty in GRACE estimates of the mass redistributions at the Earth surface: implications for the global water and sea level budgets". In: *Geophy. J. Int.* 215.1 (2018), pp. 415–430. DOI: `10.1093/gji/ggy293`.

[104] Irina Blinova, Ilya Makeev, and Igor Popov. "Benchmark solutions for stokes flows in cylindrical and spherical geometry". In: *Bulletin of the Transilvania University of Brasov. Mathematics, Informatics, Physics. Series III* 9.1 (2016), p. 11. DOI: `xxxx`.

[105] AM Bobrov and AA Baranov. "Thermochemical Mantle Convection with Drifting Deformable Continents: Main Features of Supercontinent Cycle". In: *Pure Appl. Geophys.* (2019), pp. 1–21. DOI: `10.1007/s00024-019-02164-w`.

[106] M Bocher, Nicolas Coltice, Alexandre Fournier, and Paul J Tackley. "A sequential data assimilation approach for the joint reconstruction of mantle convection and surface tectonics". In: *Geophy. J. Int.* 204.1 (2016), pp. 200–214.

[107] Marie Bocher, Alexandre Fournier, and Nicolas Coltice. "Ensemble Kalman filter for the reconstruction of the Earth's mantle circulation". In: *Nonlinear Processes in Geophysics* 25.1 (2018), pp. 99–123.

[108] P. B. Bochev, C. R. Dohrmann, and M. D. Gunzburger. "Stabilization of Low-order Mixed Finite Elements for the Stokes Equations". In: *SIAM Journal on Numerical Analysis* 44.1 (2006), pp. 82–101. DOI: `10.1137/s0036142905444482`.

[109] Pavel B Bochev and Clark R Dohrmann. "A computational study of stabilized, low-order C 0 finite element approximations of Darcy equations". In: *Computational Mechanics* 38.4-5 (2006), pp. 323–333. DOI: `10.1007/s00466-006-0036-y`.

[110] Pavel B Bochev and Max D Gunzburger. *Least-squares finite element methods*. Vol. 166. Springer Science & Business Media, 2009.

[111] Pavel B Bochev, Max D Gunzburger, and John N Shadid. "Stability of the SUPG finite element method for transient advection–diffusion problems". In: *Computer Methods in Applied Mechanics and Engineering* 193.23-26 (2004), pp. 2301–2323. DOI: `10.1016/j.cma.2004.01.026`.

[112] L. Bodri and B. Bodri. "Flow, stress and temperature in island arc areas". In: *Geophysical & Astrophysical Fluid Dynamics* 13.1 (1979), pp. 95–105. DOI: `10.1080/03091927908243763`.

[113] Reinhard Boehler. "Melting temperature of the Earth's mantle and core: Earth's thermal structure". In: *Annual Review of Earth and Planetary Sciences* 24.1 (1996), pp. 15–40.

[114] Daniele Boffi, Franco Brezzi, and Michel Fortin. "Finite elements for the Stokes problem". In: *Mixed finite elements, compatibility conditions, and applications: Lectures given at the CIME Summer School held in Cetraro, Italy, June 26-July 1, 2006.* Springer, 2008, pp. 45–100. DOI: 10.1007/978-3-540-78319-0_2.

[115] Daniele Boffi, Franco Brezzi, and Michel Fortin. *Mixed Finite Element Methods and Applications.* Springer, 2013. ISBN: 978-3-642-36518-8.

[116] Daniele Boffi, Nicola Cavallini, Francesca Gardini, and Lucia Gastaldi. "Local mass conservation of Stokes finite elements". In: *Journal of scientific computing* 52.2 (2012), pp. 383–400. DOI: 10.1007/s10915-011-9549-4.

[117] Fabian Böhm. "Matrix-free Implementation and Evaluation of the Enriched Galerkin Finite Element Method for the Stokes Problem with Varying Viscosity". MA thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg, 2023.

[118] JM Boland and RA Nicolaides. "On the stability of bilinear-constant velocity-pressure finite elements". In: *Numerische Mathematik* 44.2 (1984), pp. 219–222. DOI: 10.1007/BF01410106.

[119] JM Boland and RA Nicolaides. "Stable and semistable low order finite elements for viscous flows". In: *SIAM journal on numerical analysis* 22.3 (1985), pp. 474–492. DOI: 10.1137/0722028.

[120] Alessandro Bonaccorso and Paul M Davis. "Models of ground deformation from vertical volcanic conduits with application to eruptions of Mount St. Helens and Mount Etna". In: *J. Geophys. Res.: Solid Earth* 104.B5 (1999), pp. 10531–10542.

[121] M.-A. Bonnardot, R. Hassani, and E. Tric. "Numerical modelling of lithosphere-asthenosphere interaction in a subduction zone". In: *Earth Planet. Sci. Lett.* 272 (2008), pp. 698–708. DOI: 10.1016/j.epsl.2008.06.009.

[122] Alain Bonneville and Patrick Capolsini. "THERMIC: a 2-D finite-element tool to solve conductive and advective heat transfer problems in Earth sciences". In: *Computers & Geosciences* 25.10 (1999), pp. 1137–1148.

[123] O. Botella and R. Peyret. "Benchmark spectral results on the lid-driven cavity flow". In: *Computers and Fluids* 27.4 (1998), pp. 421–433.

[124] MHP Bott, GD Waghorn, and A Whittaker. "Plate boundary forces at subduction zones and trench-arc compression". In: *Tectonophysics* 170.1-2 (1989), pp. 1–15.

[125] Lorenzo Botti and Daniele A Di Pietro. "A pressure-correction scheme for convection-dominated incompressible flows with discontinuous velocity and continuous pressure". In: *J. Comp. Phys.* 230.3 (2011), pp. 572–585. DOI: 10.1016/j.jcp.2010.10.004.

[126] Bernard P Boudreau. *Diagenetic models and their implementation.* Vol. 505. 1997. ISBN: 978-3-642-64399-6. DOI: 10.1007/978-3-642-60421-8.

[127] Mathieu Bouffard, Stéphane Labrosse, Gaël Choblet, Alexandre Fournier, Julien Aubert, and Paul J Tackley. "A particle-in-cell method for studying double-diffusive convection in the liquid layers of planetary interiors". In: *J. Comp. Phys.* 346 (2017), pp. 552–571. DOI: 10.1016/j.jcp.2017.06.028.

[128] Edouard Boujo, Claire Bourquard, Y Xiong, and Nicolas Noiray. "Processing time-series of randomly forced self-oscillators: The example of beer bottle whistling". In: *Journal of Sound and Vibration* 464 (2020), p. 114981. DOI: 10.1016/j.jsv.2019.114981.

[129] Johannes Bouman et al. "GOCE gravitational gradients along the orbit". In: *Journal of Geodesy* 85.11 (2011), p. 791. DOI: 10.1007/s00190-011-0464-0.

[130] Johannes Bouman et al. "GOCE gravity gradient data for lithospheric modeling". In: *International Journal of Applied Earth Observation and Geoinformation* 35 (2015), pp. 16–30.

[131] L. Bourgouin, H.-B. Mühlhaus, A.J. Hale, and A. Arsac. "Studying the influence of a solid shell on lava dome growth and evolution using the level set method". In: *Geophy. J. Int.* 170 (2007), pp. 1431–1438. DOI: 10.1111/j.1365-246X.2007.03471.x.

[132] L. Bourgouin, H.-B. Mühlhaus, A.J. Hale, and A. Arsac. "Towards realistic simulations of lava dome growth using the level set method". In: *Acta Geotecnica* 1 (2006), pp. 225–236. DOI: 10.1007/s11440-006-0016-6.

[133] Allan F Bower. *Applied mechanics of solids*. CRC press, 2009. ISBN: 978-1-4398-0247-2.

[134] D.J. Bower, M. Gurnis, and D. Sun. "Dynamic origins of seismic wavespeed variation in D″". In: *Phys. Earth. Planet. Inter.* 214 (2013), pp. 74–86. DOI: 10.1016/j.pepi.2012.10.004.

[135] Malte Braack and Gert Lube. "Finite elements with local projection stabilization for incompressible flow problems". In: *Journal of Computational Mathematics* (2009), pp. 116–147. DOI: xxx.

[136] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 2007. ISBN: 978-0-521705189.

[137] D. Braess and R. Sarazin. "An Efficient Smoother for the Stokes Problem". In: *Applied Numerical Math.* 23 (1997), pp. 3–20.

[138] Henning Braess and Peter Wriggers. "Arbitrary Lagrangian Eulerian finite element analysis of free surface flow". In: *Computer Methods in Applied Mechanics and Engineering* 190.1-2 (2000), pp. 95–109. DOI: 10.1016/S0045-7825(99)00416-8.

[139] Carla Braitenberg. "Exploration of tectonic structures with GOCE in Africa and across-continents". In: *International Journal of Applied Earth Observation and Geoinformation* 35 (2015), pp. 88–95. DOI: 10.1016/j.jag.2014.01.013.

[140] James H Bramble and Joseph E Pasciak. "A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems". In: *Mathematics of Computation* 50.181 (1988), pp. 1–17.

[141] J.P. Brandenburg, E.H. Hauri, P.E. van Keken, and C.J. Ballentine. "A multiple-system study of the geochemical evolution of the mantle with force-balanced plates and thermochemical effects". In: *Earth Planet. Sci. Lett.* 276 (2008), pp. 1–13. DOI: 10.1016/j.epsl.2008.08.027.

[142] J.P. Brandenburg and P.E. van Keken. "Methods for thermochemical convection in Earth's mantle with force-balanced plates". In: *Geochem. Geophys. Geosyst.* 8.11 (2007).

[143] J. Braun. "Pecube: a new finite-element code to solve the 3D heat transport equation including the effects of a time-varying, finite amplitude surface topography". In: *Computers and Geosciences* 29 (2003), pp. 787–794. DOI: 10.1016/S0098-3004(03)00052-9.

[144] J. Braun, C. Thieulot, P. Fullsack, M. DeKool, and R.S. Huismans. "DOUAR: a new three-dimensional creeping flow model for the solution of geological problems". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 76–91. DOI: 10.1016/j.pepi.2008.05.003.

[145] J. Braun and P. Yamato. "Structural evolution of a three-dimensional, finite-width crustal wedge". In: *Tectonophysics* 484 (2010), pp. 181–192. DOI: 10.1016/j.tecto.2009.08.032.

[146] Jean Braun. "The many surface expressions of mantle dynamics". In: *Nature Geoscience* 3.12 (2010), p. 825. DOI: 10.1038/ngeo1020.

[147] Jean Braun. "Three-dimensional numerical modeling of compressional orogenies: Thrust geometry and oblique convergence". In: *Geology* 21.2 (1993), pp. 153–156. DOI: 10.1130/0091-7613(1993)021<0153:TDNMOC>2.3.CO;2.

[148] Jean Braun. "Three-dimensional numerical simulations of crustal-scale wrenching using a non-linear failure criterion". In: *Journal of Structural Geology* 16.8 (1994), pp. 1173–1186. DOI: 10.1016/0191-8141(94)90060-4.

[149] Jean Braun and Christopher Beaumont. "A physical explanation of the relation between flank uplifts and the breakup unconformity at rifted continental margins". In: *Geology* 17.8 (1989), pp. 760–764.

[150] Jean Braun and Christopher Beaumont. "Styles of continental rifting: results from dynamic models of lithospheric extension". In: *Canadian Society of Petroleum Geologists, Memoir* 12 (1987), pp. 241–258.

[151] Jean Braun and Christopher Beaumont. "Three-dimensional numerical experiments of strain partitioning at oblique plate boundaries: Implications for contrasting tectonic styles in the southern Coast Ranges, California, and central South Island, New Zealand". In: *J. Geophys. Res.: Solid Earth* 100.B9 (1995), pp. 18, 059–18, 074. DOI: 10.1029/95JB01683.

[152] Matthew C Brennan, Rebecca A Fischer, and Jessica CE Irving. "Core formation and geophysical properties of Mars". In: *Earth Planet. Sci. Lett.* 530 (2020), p. 115923. DOI: 10.1016/j.epsl.2019.115923.

[153] Susanne C Brenner. "Korn's inequalities for piecewise H1 vector fields". In: *Mathematics of Computation* (2004), pp. 1067–1087. DOI: xxxx. URL: http://www.jstor.org/stable/4099887.

[154] D Breuer, H Zhou, David A Yuen, and T Spohn. "Phase transitions in the Martian mantle: Implications for the planet's volcanic history". In: *J. Geophys. Res.: Planets* 101.E3 (1996), pp. 7531–7542. DOI: 10.1029/96JE00117.

[155] Doris Breuer, Dave A Yuen, and Tilman Spohn. "Phase transitions in the Martian mantle: Implications for partially layered convection". In: *Earth Planet. Sci. Lett.* 148.3-4 (1997), pp. 457–469. DOI: 10.1016/S0012-821X(97)00049-6.

[156] M Breuer, A Manglik, J Wicht, T Trümper, H Harder, and U Hansen. "Thermochemically driven convection in a rotating spherical shell". In: *Geophy. J. Int.* 183.1 (2010), pp. 150–162. DOI: 10.1111/j.1365-246X.2010.04722.x.

[157] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, 1991.

[158] Franco Brezzi. "On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers". In: *Publications mathématiques et informatique de Rennes* S4 (1974), pp. 1–26. DOI: xxxx.

[159] Franco Brezzi, Marie-Odile Bristeau, Leopoldo P Franca, Michel Mallet, and Gilbert Rogé. "A relationship between stabilized finite element methods and the Galerkin method with bubble functions". In: *Computer Methods in Applied Mechanics and Engineering* 96.1 (1992), pp. 117–129. DOI: 10.1016/0045-7825(92)90102-P.

[160] Franco Brezzi, Jim Douglas, Ricardo Durán, and Michel Fortin. "Mixed finite elements for second order elliptic problems in three variables". In: *Numerische Mathematik* 51 (1987), pp. 237–250. DOI: 10.1007/BF01396752.

[161] Franco Brezzi, Jim Douglas, and L. Donatella Marini. "Two families of mixed finite elements for second order elliptic problems". In: *Numerische Mathematik* 47 (1985), 217–235. DOI: 10.1007/BF01389710.

[162] Franco Brezzi, Jim Douglas, and L. Donatella Marini. "Variable degree mixed methods for second order elliptic problems". In: *Mat. Aplic. Comp.* 4.1 (1985), pp. 19–34. DOI: xxxx.

[163] Franco Brezzi, Jim Jr. Douglas, Michel Fortin, and L. Donatella Marini. "Efficient rectangular mixed finite elements in two and three space variables". In: *RAIRO: Modélisation mathématique et analyse numérique* 21 (1987), pp. 581–604. DOI: xxxx.

[164] RJ Bridwell and CA Anderson. *Thermomechanical models of the Rio Grande rift.* Tech. rep. Los Alamos Scientific Lab., NM (USA), 1980.

[165] RJ Bridwell and C Potzick. "Thermal regimes, mantle diapirs and crustal stresses of continental rifts". In: *Tectonophysics* 73.1-3 (1981), pp. 15–32.

[166] William L Briggs, Steve F McCormick, et al. *A multigrid tutorial.* Vol. 72. Siam, 2000.

[167] A.N. Brooks and T.J.R. Hughes. "Streamline Upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 32 (1982), pp. 199–259. DOI: 10.1016/0045-7825(82)90071-8.

[168] Geoffrey C Brown and Alan E Mussett. *The Inaccessible Earth: An Integrated View of its Structure and Composition, 276 pp.* Chapman and Hall, London, 1993.

[169] S. Brune and J. Autin. "The rift to break-up evolution of the Gulf of Aden: Insights from 3D numerical lithospheric-scale modelling". In: *Tectonophysics* 607 (2013), pp. 65–79. DOI: 10.1016/j.tecto.2013.06.029.

[170] C.-H. Bruneau and M. Saad. "The 2D lid-driven cavity problem revisited". In: *Computers & Fluids* 35 (2006), pp. 326–348. DOI: 10.1016/j.compfluid.2004.12.004.

[171] Sunyoung Bu, WonKyu Jung, and Philsu Kim. "An error embedded Runge-Kutta method for initial value problems". In: *Kyungpook Mathematical Journal* 56.2 (2016), pp. 311–327. DOI: 10.5666/KMJ.2016.56.2.311.

[172] W Roger Buck. "Small-scale convection induced by passive rifting: the cause for uplift of rift shoulders". In: *Earth Planet. Sci. Lett.* 77.3-4 (1986), pp. 362–372. DOI: 10.1016/0012-821X(86)90146-9.

[173] B.A. Buffet, C.W. Gable, and R.J. R.J. O'Connell. "Linear stability of a layered fluid with mobile surface plates". In: *J. Geophys. Res.: Solid Earth* 99(B10) (1994), pp. 19, 885–19, 900.

[174] H.H. Bui, R. Fukugawa, K. Sako, and S. Ohno. "Lagrangian meshfree particles method (SPH) for large deformation and failure flows of geomaterial using elastic-plastic soil constitutive model". In: *Int. J. Numer. Anal. Geomech.* 32.12 (2008), pp. 1537–1570.

[175] S. Buiter et al. "The numerical sandbox: comparison of model results for a shortening and an extension experiment". In: *Analogue and Numerical Modelling of Crustal-Scale Processes. Geological Society, London. Special Publications* 253 (2006), pp. 29–64.

[176] S.J.H. Buiter, R. Govers, and M.J.R. Wortel. "A modelling study of vertical surface displacements at convergent plate margins". In: *Geophy. J. Int.* 147 (2001), pp. 415–427. DOI: 10.1046/j.1365-246X.2001.00545.x.

[177] S.J.H. Buiter, R.S. Huismans, and C. Beaumont. "Dissipation analysis as a guide to mode selection during crustal extension and implications for the styles of sedimentary basins". In: *J. Geophys. Res.: Solid Earth* 113.B06406 (2008), B06406. DOI: 10.1029/2007JB005272.

[178] S.J.H. Buiter et al. "Benchmarking numerical models of brittle thrust wedges". In: *Journal of Structural Geology* 92 (2016), pp. 140–177. DOI: 10.1016/j.jsg.2016.03.003.

[179] A.L. Bull, M. Domeier, and T.H. Torsvik. "The effect of plate motion history on the longevity of deep mantle heterogeneities". In: *Earth Planet. Sci. Lett.* 401 (2014), pp. 172–182. DOI: 10.1016/j.epsl.2014.06.008.

[180] A.L. Bull, A.K. McNamara, T.W. Becker, and J. Ritsema. "Global scale models of the mantle flow field predicted by synthetic tomography models". In: *Phys. Earth. Planet. Inter.* 182 (2010), pp. 129–138. DOI: 10.1016/j.pepi.2010.03.004.

[181] Abigail L Bull, Allen K McNamara, and Jeroen Ritsema. "Synthetic tomography of plume clusters and thermochemical piles". In: *Earth Planet. Sci. Lett.* 278.3-4 (2009), pp. 152–162. DOI: 10.1016/j.epsl.2008.11.018.

[182] Keith Edward Bullen. *The Earth's density.* Springer Science & Business Media, 1975.

[183] P.S. Bullen. *Handbook of Means and Their Inequalities.* Springer; 2nd edition, 2003.

[184] H.-P. Bunge. "Low plume excess temperature and high core heat flux inferred from non-adiabatic geotherms in internally heated mantle circulation models". In: *Phys. Earth. Planet. Inter.* 153.1-3 (2005), pp. 3–10. DOI: 10.1016/j.pepi.2005.03.017.

[185] H.-P. Bunge, Y. Ricard, and J. Matas. "Non-adiabaticity in mantle convection". In: *Geophys. Res. Lett.* 28.5 (2001), pp. 879–882. DOI: 10.1029/2000GL011864.

[186] H.-P. Bunge, M. Richards, C. Lithgow-Bertelloni, J.R. Baumgardner, S.P. Grand, and B. Romanowicz. "Time scales and heterogeneous structure in geodynamic Earth models". In: *Science* 280 (1998), pp. 91–95.

[187] H.-P. Bunge, M.A. Richards, and J.R. Baumgardner. "A sensitivity study of three-dimensional spherical mantle convection at 10^8 Rayleigh number: Effects of depth-dependent viscosity, heating mode, and endothermic phase change". In: *J. Geophys. Res.: Solid Earth* 102.B6 (1997), pp. 11, 991–12, 007. DOI: 10.1029/96JB03806.

[188] H.-P. Bunge, M.A. Richards, and J.R. Baumgardner. "Effect of depth-dependent viscosity on the planform of mantle convection". In: *Nature* 379 (1996), pp. 436–438. DOI: 10.1038/379436a0.

[189] Martí Burcet, Beñat Oliveira, Juan Carlos Afonso, and Sergio Zlotnik. "A face-centred finite volume approach for coupled transport phenomena and fluid flow". In: *Applied Mathematical Modelling* 125 (2024), pp. 293–312. DOI: 10.1016/j.apm.2023.08.031.

[190] Guido Buresti. "A note on Stokes' hypothesis". In: *Acta Mechanica* 226 (2015), pp. 3555–3559. DOI: 10.1007/s00707-015-1380-9.

[191] J-P Burg and Yu Podladchikov. "Lithospheric scale folding: numerical modelling and application to the Himalayan syntaxes". In: *International Journal of Earth Sciences* 88.2 (1999), pp. 190–200. DOI: 10.1007/s005310050259.

[192] J.-P. Burg and S.M. Schmalholz. "Viscous heating allows thrusting to overcome crustal-scale buckling: Numerical investigation with application to the Himalayan syntaxes". In: *Earth Planet. Sci. Lett.* 274 (2008), pp. 189–203. DOI: 10.1016/j.epsl.2008.07.022.

[193] Roland Bürgmann and Georg Dresen. "Rheology of the lower crust and upper mantle: Evidence from rock mechanics, geodesy, and field observations". In: *Annu. Rev. Earth Planet. Sci.* 36 (2008), pp. 531–567. DOI: 10.1146/annurev.earth.36.031207.124326.

[194] Gilmer R Burgos, Andreas N Alexandrou, and Vladimir Entov. "On the determination of yield surfaces in Herschel–Bulkley fluids". In: *Journal of Rheology* 43.3 (1999), pp. 463–483. DOI: 10.1122/1.550992.

[195] E. Burman and P. Hansbo. "Edge stabilization for the generalized Stokes problem: A continuous interior penalty method". In: *Computer Methods in Applied Mechanics and Engineering* 195 (2006), pp. 2393–2410. DOI: 10.1016/j.cma.2005.05.009.

[196] Erik Burman and Peter Hansbo. "A unified stabilized method for Stokes' and Darcy's equations". In: *Journal of Computational and Applied Mathematics* 198.1 (2007), pp. 35–51. DOI: 10.1016/j.cam.2005.11.022.

[197] E. Burov and S. Cloetingh. "Erosion and rift dynamics: new thermomechanical aspects of post-rift evolution of extensional basins". In: *Earth Planet. Sci. Lett.* 150 (1997), pp. 7–26. DOI: 10.1016/S0012-821X(97)00069-1.

[198] E. Burov and A. Poliakov. "Erosion and rheology controls on synrift and postrift evolution: Verifying old and new ideas using a fully coupled numerical model". In: *J. Geophys. Res.: Solid Earth* 106.B8 (2001), pp. 16, 461–16, 481. DOI: 10.1029/2001JB000433.

[199] EB Burov and P Molnar. "Small and large-amplitude gravitational instability of an elastically compressible viscoelastic Maxwell solid overlying an inviscid incompressible fluid: dependence of growth rates on wave number and elastic constants at low Deborah numbers". In: *Earth Planet. Sci. Lett.* 275.3-4 (2008), pp. 370–381. DOI: 10.1016/j.epsl.2008.08.032.

[200] EB Burov, AB Watts, et al. "The long-term strength of continental lithosphere: "jelly sandwich" or "crème brûlée"?" In: *GSA today* 16.1 (2006), p. 4. DOI: 10.1130/1052-5173(2006)016<4:TLTSOC>2.0.CO;2.

[201] Evgene B Burov. "Rheology and strength of the lithosphere". In: *Marine and Petroleum Geology* 28.8 (2011), pp. 1402–1443. DOI: 10.1016/j.marpetgeo.2011.05.008.

[202] C. Burstedde, O. Ghattas, G. Stadler, T. Tu, and L.C. Wilcox. "Parallel scalable adjoint-based adaptive solution of variable-viscosity Stokes flow problems". In: *Computer Methods in Applied Mechanics and Engineering* 198 (2009), pp. 1691–1700. DOI: 10.1016/j.cma.2008.12.015.

[203] C. Burstedde et al. "Large-scale adaptive mantle convection simulation". In: *Geophy. J. Int.* 192 (2013), pp. 889–906. DOI: 10.1093/gji/ggs070.

[204] C. Burstedde et al. "Scalable Adaptive Mantle Convection Simulation on Petascale Supercomputers". In: *ACM/IEEE SC Conference Series, 2008* (2008).

[205] Carsten Burstedde, Lucas C. Wilcox, and Omar Ghattas. "p4est: Scalable Algorithms for Parallel Adaptive Mesh Refinement on Forests of Octrees". In: *SIAM Journal on Scientific Computing* 33.3 (2011), pp. 1103–1133. DOI: 10.1137/100791634.

[206] Carsten Burstedde et al. "Extreme-scale AMR". In: *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis.* IEEE Computer Society. 2010, pp. 1–12. DOI: 10.1109/SC.2010.25.

[207] F.H. Busse et al. "3D convection at infinite Prandtl number in Cartesian geometry - a benchmark comparison". In: *Geophysical & Astrophysical Fluid Dynamics* 75.1 (1994), pp. 39–59. DOI: 10.1080/03091929408203646.

[208] Friedrich H Busse. "Fundamentals of thermal convection". In: (1989).

[209] J. Butcher. *Numerical Methods for Ordinary Differential Equations.* New York: John Wiley & Sons, 2003.

[210] Samuel L Butler and Gary T Jarvis. "Stresses induced in continental lithospheres by axisymmetric spherical convection". In: *Geophy. J. Int.* 157.3 (2004), pp. 1359–1376. DOI: 10.1111/j.1365-246X.2004.02257.x.

[211] James Byerlee. "Friction of rocks". In: *Rock friction and earthquake prediction.* 1978, pp. 615–626.

[212] Mauro Cacace and Antoine B Jacquey. "Flexible parallel implicit modelling of coupled thermal–hydraulic–mechanical processes in fractured rocks". In: *Solid Earth* 8.5 (2017), pp. 921–941. DOI: 10.5194/se-2017-33.

[213] O. Cadek and D.A. Yuen. "GLOBAL GEODYNAMICS: Geodynamics workshop in the Czech Republic". In: *Terra Nova* 5.6 (1993), pp. 573–590. DOI: 10.1111/j.1365-3121.1993.tb00308.x.

[214] O Čadek and AP van den Berg. "Radial profiles of temperature and viscosity in the Earth's mantle inferred from the geoid and lateral seismic structure". In: *Earth Planet. Sci. Lett.* 164.3-4 (1998), pp. 607–615. DOI: 10.1016/S0012-821X(98)00244-1.

[215] C Cadio, I Panet, A Davaille, M Diament, L Métivier, and O de Viron. "Pacific geoid anomalies revisited in light of thermochemical oscillating domes in the lower mantle". In: *Earth Planet. Sci. Lett.* 306.1-2 (2011), pp. 123–135. DOI: 10.1016/j.epsl.2011.03.040.

[216] Neil Cagney, William H Newsome, Carolina Lithgow-Bertelloni, Aline Cotel, Stanley R Hart, and John A Whitehead. "Temperature and velocity measurements of a rising thermal plume". In: *Geochem. Geophys. Geosyst.* 16.3 (2015), pp. 579–599. DOI: 10.1002/2014GC005576.

[217] J. Cahouet and J.-P. Chabard. "Some fast 3D finite element solvers for the generalized Stokes problem". In: *Int. J. Num. Meth. Fluids* 8 (1988), pp. 869–895.

[218] Z Cai, Jim Douglas Jr, Juan E Santos, Dongwoo Sheen, and Xiu Ye. "Nonconforming quadrilateral finite elements: a correction". In: *Calcolo* 37.4 (2000), pp. 253–254. DOI: 10.1007/s100920070004.

[219] Zhiqiang Cai, Jim Douglas, and Xiu Ye. "A stable nonconforming quadrilateral finite element method for the stationary Stokes and Navier–Stokes equations". In: *Calcolo* 36.4 (1999), pp. 215–232. DOI: 10.1007/s100920050031.

[220] Marco Camesasca, Miron Kaufman, and Ica Manas-Zloczower. "Quantifying fluid mixing with the Shannon entropy". In: *Macromolecular theory and simulations* 15.8 (2006), pp. 595–607. DOI: 10.1002/mats.200600037.

[221] Lorenzo G Candioti, Stefan M Schmalholz, and Thibault Duretz. "Impact of upper mantle convection on lithosphere hyperextension and subsequent horizontally forced subduction initiation". In: *Solid Earth* 11.6 (2020), pp. 2327–2357. DOI: 10.5194/se-11-2327-2020.

[222] T-S Cao, Pierre Montmitonnet, and P-O Bouchard. "A detailed description of the Gurson–Tvergaard–Needleman model within a mixed velocity–pressure finite element formulation". In: *Int. J. Num. Meth. Eng.* 96.9 (2013), pp. 561–583. DOI: 10.1002/nme.4571.

[223] Z.-H. Cao. "Fast Uzawa algorithm for generalized saddle point problems". In: *Applied Numerical Mathematics* 46 (2003), pp. 157–171.

[224] G.F. Carey and J.T. Oden. *Finite Elements: Fluid Mechanics. Vol. VI.* Englewood Cliffs, Prentice-Hall, 1986. ISBN: 0-13-317132-9.

[225] P Carré, AJF Metherell, and TJ Quinn. "Apropos of "The Gravitational Field of a 111 Tetrahedron"". In: *Metrologia* 23.2 (1986), p. 119. DOI: 10.1088/0026-1394/23/2/007.

[226] P.J. Carreau. "Rheological Equations from Molecular Network Theories". In: *Transactions of the Society of rheology* 16.1 (1972), pp. 99–127. DOI: 10.1122/1.549276.

[227] J. Carrero, B. Cockburn, and D. Schötzau. "Hybridized globally divergence-free LDG methods. Part I: The Stokes problem". In: *Mathematics of Computation* 75.254 (2005), pp. 533–563.

[228] Carsten Carstensen, Karonline Köhler, Daniel Peterseim, and Mira Schedensack. "Comparison results for the Stokes equations". In: *Applied Numerical Mathematics* 95 (2015), pp. 118–129. DOI: 10.1016/j.apnum.2013.12.005.

[229] Michael A Case, Vincent J Ervin, Alexander Linke, and Leo G Rebholz. "A connection between Scott–Vogelius and grad-div stabilized Taylor–Hood FE approximations of the Navier–Stokes equations". In: *SIAM Journal on Numerical Analysis* 49.4 (2011), pp. 1461–1481. DOI: 10.1137/100794250.

[230] J.R. Cash and A.H. Karp. "A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides". In: *ACM Transcations on Mathematical Software* 16.3 (1990), pp. 201–222.

[231] Patrick Cassen and Ray T Reynolds. "Convection in the Moon: Effect of variable viscosity". In: *J. Geophys. Res.: Solid Earth* 79.20 (1974), pp. 2937–2944. DOI: 10.1029/JB079i020p02937.

[232] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau. "Local discontinuous Galerkin methods for elliptic problems". In: *Commun. Numer. Meth. Engng* 18 (2002), pp. 69–75.

[233] N.G. Cerpa, R. Hassani, M. Gerbault, and J.-H. Prévost. "A fictitious domain method for lithosphere-asthenosphere interaction: Application to periodic slab folding in the upper mantle". In: *Geochem. Geophys. Geosyst.* 15.5 (2014), pp. 1852–1877. DOI: 10.1002/2014GC005241.

[234] Jagabanduhu Chakrabarty. *Theory of plasticity.* 2012. ISBN: 0-7506-6638-2.

[235] TGG Chanut, Safwan Aljbaae, and V Carruba. "Mascon gravitation model using a shaped polyhedral source". In: *Monthly Notices of the Royal Astronomical Society* 450.4 (2015), pp. 3742–3749. DOI: 10.1093/mnras/stv845.

[236] D.S. Chapman. "Thermal gradients in the continental crust". In: *Geological Society, London, Special Publications* 24 (1986), pp. 63–70. DOI: 10.1144/GSL.SP.1986.024.01.07.

[237] D. Chappelle and K.J. Bathe. "The Inf-Sup test". In: *Computers & Structures* 47.4/5 (1993), pp. 537–545.

[238] J.S. Chen, C. Pan, and T.Y.P. Chang. "On the control of pressure oscillation in bilinear-displacement constant-pressure element". In: *Computer Methods in Applied Mechanics and Engineering* 128 (1995), pp. 137–152.

[239] Long Chen. "A simple construction of a Fortin operator for the two dimensional Taylor–Hood element". In: *Computers & Mathematics with Applications* 68.10 (2014), pp. 1368–1373. DOI: 10.1016/j.camwa.2014.09.003.

[240] Qingshan Chen, Max Gunzburger, and Mauro Perego. "Well-posedness results for a nonlinear Stokes problem arising in glaciology". In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 2710–2733. DOI: 10.1137/110848694.

[241] Z. Chen. "Analysis of mixed methods using conforming and nonconforming finite element methods". In: *Modélisation mathématique et analyse numérique* 27.1 (1993), pp. 9–34. DOI: xxxx.

[242] Z. Chen. *Finite Element Methods and Their Applications.* Springer, 2005.

[243] Z. Chen. "Projection Finite Element methods for semiconductor device equations". In: *Computers Math. Applic.* 25.8 (1993), pp. 81–88. DOI: 10.1016/0898-1221(93)90173-S.

[244] Z. Chen and P. Oswald. "Multigrid and multilevel methods for nonconforming Q_1 elements". In: *Mathematic of Computation* 67.222 (1998), pp. 667–693.

[245] P. Chenin and C. Beaumont. "Influence of offset weak zones on the development of rift basins: Activation and abandonment during continental extension and breakup". In: *J. Geophys. Res.: Solid Earth* 118 (2013), pp. 1–23. DOI: 10.1002/jgrb.50138.

[246] P. Chenin, S.M. Schmalholz, G. Manatschal, and G.D. Karner. "Necking of the Lithosphere: A Reappraisal of Basic Concepts With Thermo-Mechanical Numerical Modeling". In: *J. Geophys. Res.: Solid Earth* 123.6 (2018), pp. 5279–5299. DOI: 10.1029/2017JB014155.

[247] Pauline Chenin, Gianreto Manatschal, Alessandro Decarlis, Stefan M Schmalholz, Thibault Duretz, and Marco Beltrando. "Emersion of distal domains in advanced stages of continental rifting explained by asynchronous crust and mantle necking". In: *Geochem. Geophys. Geosyst.* (2019). DOI: 10.1029/2019GC008357.

[248] Pauline Chenin, Stefan M Schmalholz, Gianreto Manatschal, and Thibault Duretz. "Impact of crust–mantle mechanical coupling on the topographic and thermal evolutions during the necking phase of 'magma-poor' and 'sediment-starved' rift systems: A numerical modeling study". In: *Tectonophysics* (2020), p. 228472. DOI: 10.1016/j.tecto.2020.228472.

[249] M.V. Chertova, T. Geenen, A. van den Berg, and W. Spakman. "Using open sidewalls for modelling self-consistent lithosphere subduction dynamics". In: *Solid Earth* 3 (2012), pp. 313–326. DOI: 10.5194/se-3-313-2012.

[250] Jean Chery, Francis Lucazeau, Marc Daignieres, and Jean-Pierre Vilotte. "Large uplift of rift flanks: A genetic link with lithospheric rigidity?" In: *Earth Planet. Sci. Lett.* 112.1-4 (1992), pp. 195–211. DOI: 10.1016/0012-821X(92)90016-O.

[251] Alima Chibani and Nasserdine Kechkar. "Minimal Locally Stabilized Q1-Q0 Schemes for the Generalized Stokes Problem". In: *Journal of the Korean Mathematical Society* 57.5 (2020), pp. 1239–1266. DOI: 10.4134/JKMS.j190628.

[252] S. Chiu-Webster, E.J. Hinch, and J.R. Lister. "Very viscous horizontal convection". In: *J. Fluid Mech.* 611 (2008), pp. 395–426.

[253] G. Choblet. "Modelling thermal convection with large viscosity gradients in one block of the 'cubed sphere'". In: *J. Comp. Phys.* 205 (2005), pp. 269–291.

[254] G. Choblet, O. Čadek, F. Couturier, and C. Dumoulin. "OEDIPUS: a new tool to study the dynamics of planetary interiors". In: *Geophy. J. Int.* 170 (2007), pp. 9–30.

[255] E. Choi and K.D. Petersen. "Making Coulomb angle-oriented shear bands in numerical tectonic models". In: *Tectonophysics* 657 (2015), pp. 94–101. DOI: 10.1016/j.tecto.2015.06.026.

[256] E. Choi, E. Tan, L.L. Lavier, and V.M. Calo. "DynEarthSol2D: An efficient unstructured finite element method to study long-term tectonic deformation". In: *J. Geophys. Res.: Solid Earth* 118 (2013), pp. 1–16. DOI: 10.1002/jgrb.50148.

[257] A.J. Chorin. "A numerical method for solving incompressible viscous flow problems". In: *J. Comp. Phys.* 2 (1967), pp. 12–26. DOI: 10.1006/jcph.1997.5716.

[258] Nikolas I Christensen and Walter D Mooney. "Seismic velocity structure and composition of the continental crust: A global view". In: *J. Geophys. Res.: Solid Earth* 100.B6 (1995), pp. 9761–9788. DOI: /10.1029/95JB00259.

[259] R.R. Christensen. "An Eulerian technique for thermomechanical modeling of lithospheric extension". In: *J. Geophys. Res.: Solid Earth* 97.B2 (1992), pp. 2015–2036. DOI: 10.1029/91JB02642.

[260] U Christensen. "Convection in a variable-viscosity fluid: Newtonian versus power-law rheology". In: *Earth Planet. Sci. Lett.* 64.1 (1983), pp. 153–162. DOI: 10.1016/0012-821X(83)90060-2.

[261] U Christensen. "Convection with pressure-and temperature-dependent non-Newtonian rheology". In: *Geophy. J. Int.* 77.2 (1984), pp. 343–384. DOI: 10.1111/j.1365-246X.1984.tb01939.x.

[262] U Christensen. "Mixing by time-dependent convection". In: *Earth Planet. Sci. Lett.* 95.3-4 (1989), pp. 382–394. DOI: 10.1016/0012-821X(89)90112-X.

[263] U.R. Christensen and D.A. Yuen. "The interaction of a subducting lithospheric slab with a chemical or phase boundary". In: *J. Geophys. Res.: Solid Earth* 89(B6) (1984), pp. 4389–4402. DOI: 10.1029/JB089iB06p04389.

[264] U.R. Christensen and D.A. Yuen. "Time-dependent convection with non-Newtonian viscosity". In: *J. Geophys. Res.: Solid Earth* 94.B1 (1989), pp. 814–820. DOI: 10.1029/JB094iB01p00814.

[265] Ulrich Christensen. "A numerical model of coupled subcontinental and oceanic convection". In: *Tectonophysics* 95.1-2 (1983), pp. 1–23. DOI: 10.1016/0040-1951(83)90256-1.

[266] Ulrich Christensen. "Phase boundaries in finite amplitude mantle convection". In: *Geophy. J. Int.* 68.2 (1982), pp. 487–497. DOI: 10.1111/j.1365-246X.1982.tb04911.x.

[267] Ulrich R Christensen. "The influence of trench migration on slab penetration into the lower mantle". In: *Earth Planet. Sci. Lett.* 140.1-4 (1996), pp. 27–39. DOI: 10.1016/0012-821X(96)00023-4.

[268] Ulrich R Christensen. "Time-dependent convection in elongated Rayleigh-Benard cells". In: *Geophys. Res. Lett.* 14.3 (1987), pp. 220–223. DOI: 10.1029/GL014i003p00220.

[269] Ulrich R Christensen and David A Yuen. "Layered convection induced by phase transitions". In: *J. Geophys. Res.: Solid Earth* 90.B12 (1985), pp. 10291–10300. DOI: 10.1029/JB090iB12p10291.

[270] Edmund Christiansen and Knud D. Andersen. "Computation of collapse states with von Mises type yield condition". In: *Int. J. Num. Meth. Eng.* 46 (1999), pp. 1185–1202.

[271] Edmund Christiansen and Ole S. Pedersen. "Automatic mesh refinement in limit analysis". In: *Int. J. Num. Meth. Eng.* 50 (2001), pp. 1331–1346.

[272] M.A. Christon. "Dealing with pressure: FEM solution strategies for the pressure in the time-dependent Navier-Stokes equations". In: *Int. J. Num. Meth. Fluids* 38 (2002), pp. 1177–1198.

[273] M.A. Christon and G.O. Cook. *LS-DYNA's Incompressible Flow Solver User's Manual.* Livermore Software Technology Corporation.

[274] M.A. Christon, P.M. Gresho, and S.B. Sutton. "Computational predictability of time-dependent natural convection flows in enclosures (including a benchmark solution)". In: *Int. J. Num. Meth. Fluids* 40 (2002), pp. 953–980.

[275] Mark A Christon. "The new incompressible flow capabilities in LS-DYNA". In: *6th International LS-DYNA Users Conference.* 2000.

[276] Andrea Cioncolini and Daniele Boffi. "The MINI mixed finite element for the Stokes problem: An experimental investigation". In: *Computers & Mathematics with Applications* 77.9 (2019), pp. 2432–2446. DOI: 10.1016/j.camwa.2018.12.028.

[277] H. Čížková, A.P. van den Berg, W. Spakman, and Ctirad Matyska. "The viscosity of the earth's lower mantle inferred from sinking speed of subducted lithosphere". In: *Phys. Earth. Planet. Inter.* 200–201 (2012), pp. 56–62. DOI: 10.1016/j.pepi.2012.02.010.

[278] T.C. Clevenger, T. Heister, G. Kanschat, and M. Kronbichler. "A flexible, parallel, adaptive geometric multigrid method for FEM". In: *ACM Trans. Math. Softw.* 47.1 (2020), pp. 1–27. DOI: 10.1145/3425193.

[279] Thomas C Clevenger and Timo Heister. "Comparison Between Algebraic and Matrix-free Geometric Multigrid for a Stokes Problem on Adaptive Meshes with Variable Viscosity". In: *Numer. Linear Algebra Appl.* (2021). DOI: 10.1002/nla.2375.

[280] S Cochard and C Ancey. "Experimental investigation of the spreading of viscoplastic fluids on inclined planes". In: *Journal of Non-Newtonian Fluid Mechanics* 158.1-3 (2009), pp. 73–84.

[281] B. Cockburn and J. Gopalakrishnan. "The derivation of hybridizable discontinuous Galerkin methods for Stokes flow". In: *SIAM J. Numer. Anal.* 47.2 (2009), pp. 1092–1125.

[282] B. Cockburn, G. Kanschat, and D. Schötzau. "The local discontinuous Galerkin method for linearized incompressible fluid flow: a review". In: *Computers and Fluids* 34 (2005), pp. 491–506.

[283] B. Cockburn, G. Kansschat, D. Schoetzau, and C. Schwab. "Local discontinuous Galerkin methods for the Stokes system". In: *SIAM J. Numer. Anal.* 40.1 (2002), pp. 319–343.

[284] B. Cockburn, G.E. Karniadakis, and C.W.Shu. "The Development of Discontinuous Galerkin Methods". In: *Discontinuous Galerkin Methods. Lecture Notes in Computational Science and Engineering* 11 (2000).

[285] B. Cockburn, G.E. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin Methods. Theory, Computation and Applications.* Springer, 2000.

[286] B. Cockburn, N.C. Nguyen, and J. Peraire. "A Comparison of HDG Methods for Stokes Flow". In: *J. Sci. Comput.* 45 (2010), pp. 215–237.

[287] Bernardo Cockburn, Guido Kanschat, and Dominik Schötzau. "The local discontinuous Galerkin method for the Oseen equations". In: *Mathematics of Computation* 73.246 (2004), pp. 569–593.

[288] Ramon Codina. "On stabilized finite element methods for linear systems of convection–diffusion-reaction equations". In: *Computer Methods in Applied Mechanics and Engineering* 188.1-3 (2000), pp. 61–82. DOI: 10.1016/S0045-7825(00)00177-8.

[289] L. Colli, H.-P. Bunge, and B.S.A. Schuberth. "On retrodictions of global mantle flow with assimilated surface velocities". In: *Geophys. Res. Lett.* 42.20 (2015), pp. 8341–8348. DOI: 10.1002/2015GL066001.

[290] N. Coltice. "The role of convective mixing in degassing the Earth's mantle". In: *Earth Planet. Sci. Lett.* 234.1-2 (2005), pp. 15–25. DOI: 10.1016/j.epsl.2005.02.041.

[291] Nicolas Coltice, Laurent Husson, Claudio Faccenna, and Maëlis Arnould. "What drives tectonic plates?" In: *Science Advances* 5.10 (2019). DOI: 10.1126/sciadv.aax4295.

[292] Nicolas Coltice and J Schmalzl. "Mixing times in the mantle of the early Earth derived from 2-D and 3-D numerical simulations of convection". In: *Geophys. Res. Lett.* 33.23 (2006). DOI: 10.1029/2006GL027707.

[293] Nicolas Coltice and Grace E Shephard. "Tectonic predictions with mantle convection models". In: *Geophy. J. Int.* 213.1 (2018), pp. 16–29. DOI: 10.1093/gji/ggx531.

[294] G Comini, Marco Manzan, and C Nonino. "Finite element solution of the streamfunction-vorticity equations for incompressible two-dimensional flows". In: *Int. J. Num. Meth. Fluids* 19.6 (1994), pp. 513–525. DOI: 10.1002/fld.1650190605.

[295] James AD Connolly and Yury Y Podladchikov. "An analytical solution for solitary porosity waves: dynamic permeability and fluidization of nonlinear viscous and viscoplastic rock". In: *Geofluids* 15.1-2 (2015), pp. 269–292. DOI: `10.1111/gfl.12110`.

[296] Clinton P Conrad and Peter Molnar. "The growth of Rayleigh-Taylor-type instabilities in the lithosphere for various rheological and density structures". In: *Geophy. J. Int.* 129.1 (1997), pp. 95–112.

[297] CM Cooper, A Lenardic, A Levander, L Moresi, and K Benn. "Creation and preservation of cratonic lithosphere: seismic constraints and geodynamic models". In: *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION* 164 (2006), p. 75. DOI: `xxxx`.

[298] C. Couder-Castaneda, J.C. Ortiz-Aleman, M.G. Orozco-del-Castillo, and M. Nava-Flores. "Forward modeling of gravitational fields on hybrid multi-threaded cluster". In: *Geofisica Internacional* 54.1 (2015), pp. 31–48.

[299] D Coumou, T Driesner, and Christoph A Heinrich. "The structure and dynamics of mid-ocean ridge hydrothermal systems". In: *Science* 321.5897 (2008), pp. 1825–1828.

[300] P. Coussot. "Yield stress fluid flows: A review of experimental data". In: *Journal of Non-Newtonian Fluid Mechanics* 211 (2014), pp. 31–49.

[301] Claire Harvey Craig and Dan McKenzie. "The existence of a thin low-viscosity layer beneath the lithosphere". In: *Earth Planet. Sci. Lett.* 78.4 (1986), pp. 420–426.

[302] F. Crameri and P.J. Tackley. "Spontaneous development of arcuate single-sided subduction in global 3-D mantle convection models with a free surface". In: *J. Geophys. Res.: Solid Earth* 119 (2014). DOI: `10.1002/2014JB010939`.

[303] F. Crameri and P.J. Tackley. "Subduction initiation from a stagnant lid and global overturn: new insights from numerical models with a free surface". In: *Progress in Earth and Planetary Science* 3 (2016).

[304] F. Crameri et al. "A comparison of numerical surface topography calculations in geodynamic modelling: an evaluation of the 'sticky air' method". In: *Geophy. J. Int.* 189 (2012), pp. 38–54. DOI: `10.1111/j.1365-246X.2012.05388.x`.

[305] Fabio Crameri, Grace E. Shephard, and Philip J. Heron. "The misuse of colour in science communication". In: *Nature Communications* 11 (2020), p. 5444. DOI: `10.1038/s41467-020-19160-7`.

[306] Marc Crombaghs, Erik de Min, and Govert Strang van Hees. "The first absolute gravity measurements in The Netherlands". In: *TU Delft, Delft, Tech. Rep* (2002).

[307] M.M. Cross. "Rheology of non-Newtonian fluids: a new flow equation for pseudoplastic systems". In: *Journal of Colloid Science* 20 (1965), pp. 417–437.

[308] M. Crouzeix and R.S. Falk. "Nonconforming finite elements for the Stokes problem". In: *Mathematic of Computation* 52.186 (1989), pp. 437–456. DOI: `10.1090/S0025-5718-1989-0958870-8`.

[309] M. Crouzeix and P.-A. Raviart. "Conforming and nonconforming finite element methods for solving the stationary Stokes equations I". In: *R.A.I.R.O.* 7.3 (1973), pp. 33–75. DOI: `10.1051/m2an/197307R300331`. URL: `http://eudml.org/doc/193250`.

[310] AR Cruden. "Deformation around a rising diapir modeled by creeping flow past a sphere". In: *Tectonics* 7.5 (1988), pp. 1091–1101.

[311] L. Cserepes and D.A. Yuen. "Dynamical consequences of mid-mantle viscosity stratification on mantle flows with an endothermic phase transition". In: *Geophys. Res. Lett.* 24.2 (1997), pp. 181–184. DOI: `10.1029/96GL03917`.

[312] Marco Cuffaro, Edie Miglio, Mattia Penati, and Marco Viganò. "Mantle thermal structure at northern Mid-Atlantic Ridge from improved numerical methods and boundary conditions". In: *Geophy. J. Int.* 220.2 (2020), pp. 1128–1148. DOI: 10.1093/gji/ggz488.

[313] Gilda Currenti and Charles A Williams. "Numerical modeling of deformation and stress fields around a magma chamber: Constraints on failure conditions and rheology". In: *Phys. Earth. Planet. Inter.* 226 (2014), pp. 14–27. DOI: 10.1016/j.pepi.2013.11.003.

[314] C.A. Currie and C. Beaumont. "Are diamond-bearing Cretaceous kimberlites related to low-angle subduction beneath western North America". In: *Earth Planet. Sci. Lett.* 303 (2011), pp. 59–70. DOI: 10.1016/j.epsl.2010.12.036.

[315] Claire A Currie and Roy D Hyndman. "The thermal structure of subduction zone back arcs". In: *J. Geophys. Res.: Solid Earth* 111.B8 (2006). DOI: 10.1029/2005JB004024.

[316] Elizabeth Cuthill and James McKee. "Reducing the bandwidth of sparse symmetric matrices". In: *Proceedings of the 1969 24th national conference.* 1969, pp. 157–172. DOI: 10.1145/800195.805928.

[317] C. Cuvelier, A. Segal, and A.A. van Steenhoven. *Finite Element Methods and Navier-Stokes Equations.* D. Reidel Publishing Company, 1986.

[318] M. Dabrowski, M. Krotkiewski, and D.W. Schmid. "MILAMIN: Matlab based finite element solver for large problems". In: *Geochem. Geophys. Geosyst.* 9.4 (2008), Q04030. DOI: 10.1029/2007GC001719.

[319] S.F. Daly and A. Raefsky. "On the penetration of a hot diapir through a strongly temperature-dependent viscosity medium". In: *Geophys. J. R. astr. Soc* 83 (1985), pp. 657–681. DOI: 10.1111/j.1365-246X.1985.tb04331.x.

[320] J. Dannberg, Z. Eilon, U. Faul, R. Gassmöller, P. Moulik, and R. Myhill. "The importance of grain size to mantle dynamics and seismological observations". In: *Geochem. Geophys. Geosyst.* 18 (2017), pp. 3034–3061. DOI: 10.1002/2017GC006944.

[321] J. Dannberg and T. Heister. "Compressible magma/mantle dynamics: 3-D, adaptive simulations in ASPECT". In: *Geophy. J. Int.* 207 (2016), pp. 1343–1366. DOI: 10.1093/gji/ggw329.

[322] Juliane Dannberg, Rene Gassmöller, Ryan Grove, and Timo Heister. "A new formulation for coupled magma/mantle dynamics". In: *Geophy. J. Int.* 219.1 (2019), pp. 94–107. DOI: 10.1093/gji/ggz190.

[323] KD Danov, TD Gurkov, H Raszillier, and F Durst. "Stokes flow caused by the motion of a rigid sphere close to a viscous interface". In: *Chemical Engineering Science* 53.19 (1998), pp. 3413–3434.

[324] Véronique Dansereau, Jérôme Weiss, Pierre Saramito, and Philippe Lattes. "A Maxwell-Elasto-Brittle rheology for sea ice modelling". In: *The Cryosphere* 10 (2016), pp. 1339–1359. DOI: 10.5194/tc-10-1339-2016.

[325] D Rhodri Davies, Stephan C Kramer, Sia Ghelichkhan, and Angus Gibson. "Towards automatic finite-element methods for geodynamics via Firedrake". In: *Geosci. Model. Dev.* 15.13 (2022), pp. 5127–5166. DOI: 10.5194/gmd-15-5127-2022.

[326] D. R. Davies, J. H. Davies, O. Hassan, K. Morgan, and P. Nithiarasu. "Investigations into the applicability of adaptive finite element methods to two-dimensional infinite Prandtl number thermal and thermochemical convection". In: *Geochem. Geophys. Geosyst.* 8.5 (2007). DOI: 10.1029/2006GC001470.

[327] D.R. Davies, J.H. Davies, P.C. Bollada, O. Hassan, K. Morgan, and P. Nithiarasu. "A hierarchical mesh refinement technique for global 3-D spherical mantle convection modelling". In: *Geosci. Model. Dev.* 6 (2013), pp. 1095–1107. DOI: 10.5194/gmd-6-1095-2013.

[328] D.R. Davies, C.R. Wilson, and S.C. Kramer. "Fluidity: A fully unstructured anisotropic adaptive mesh computational modeling framework for geodynamics". In: *Geochem. Geophys. Geosyst.* 12.6 (2011), Q06001. DOI: 10.1029/2011GC003551.

[329] David Rhodri Davies, John Huw Davies, O Hassan, K Morgan, and P Nithiarasu. "Adaptive finite element methods in geodynamics: Convection dominated mid-ocean ridge and subduction zone simulations". In: *International Journal of Numerical Methods for Heat & Fluid Flow* 18.7/8 (2008). DOI: 10.1108/09615530810899079.

[330] G.F. Davies. "Mantle convection model with a dynamic plate: topography, heat flow and gravity anomalies". In: *Geophy. J. Int.* 98.3 (1989), pp. 461–464. DOI: 10.1111/j.1365-246X.1989.tb02283.x.

[331] G.F. Davies. "Role of the lithosphere in mantle convection". In: *J. Geophys. Res.: Solid Earth* 93.B9 (1988), pp. 10451–10466. DOI: 10.1029/JB093iB09p10451.

[332] Geoffrey F Davies. "Mantle convection under simulated plates: effects of heating modes and ridge and trench migration, and implications for the core-mantle boundary, bathymetry, the geoid and Benioff zones". In: *Geophys. J. R. astr. Soc* 84.1 (1986), pp. 153–183.

[333] Geoffrey F Davies. "Viscous mantle flow under moving lithospheric plates and under subduction zones". In: *Geophy. J. Int.* 49.3 (1977), pp. 557–563. DOI: 10.1111/j.1365-246X.1977.tb01303.x.

[334] Geoffrey F Davies. "Whole-mantle convection and plate tectonics". In: *Geophy. J. Int.* 49.2 (1977), pp. 459–486. DOI: 10.1111/j.1365-246X.1977.tb03717.x.

[335] Geoffrey F Davies and Michael Gurnis. "Interaction of mantle dregs with convection: Lateral heterogeneity at the core-mantle boundary". In: *Geophys. Res. Lett.* 13.13 (1986), pp. 1517–1520.

[336] J. H. Davies and D. J. Stevenson. "Physical model of source region of subduction zone volcanics". In: *J. Geophys. Res.: Solid Earth* 97.B2 (1992), pp. 2037–2070. DOI: 10.1029/91JB02571.

[337] R.O. Davis and A.P.S. Selvadurai. *Elasticity and Geomechanics*. Cambridge University Press, 1996.

[338] R.O. Davis and A.P.S. Selvadurai. *Plasticity and Geomechanics*. Cambridge University Press, 2002. ISBN: 0-521-01809-9.

[339] P. Davy and P. Cobbold. "Indentation tectonics in nature and experiment. 1. Experiments scaled for gravity". In: *Bulletin of the Geological Institutions of Uppsala* 14 (1988), pp. 129–141.

[340] Arne De Coninck et al. "Needles: Toward Large-Scale Genomic Prediction with Marker-by-Environment Interaction". In: *Genetics* 203.1 (2016), pp. 543–555. DOI: 10.1534/genetics.115.179887.

[341] Albert de Montserrat, Jason P Morgan, and Jörg Hasenclever. "LaCoDe: a Lagrangian two-dimensional thermo-mechanical code for large-strain compressible visco-elastic geodynamical modeling". In: *Tectonophysics* 767 (2019), p. 228173. DOI: 10.1016/j.tecto.2019.228173.

[342] J.H. De Smet, Arie P van Den Berg, Nico J Vlaar, and David A Yuen. "A characteristics-based method for solving the transport equation and its application to the process of mantle differentiation and continental root growth". In: *Geophy. J. Int.* 140.3 (2000), pp. 651–659. DOI: 10.1046/j.1365-246X.2000.00993.x.

[343] A. Deb, J.H. Prevost, and B. Loret. "Adaptive meshing for dynamic strain localisation". In: *Comput. Methods Appl. Mech. Engrg.* 137 (1996), pp. 285–306. DOI: 10.1016/S0045-7825(96)01068-7.

[344] Jembere Degefa. "Numerical solution of the two-dimensional wave equation using Crank-Nicolson scheme". MA thesis. Jimma University, 2018.

[345] Herwig Dejonghe. "A completely analytical family of anisotropic Plummer models". In: *Monthly Notices of the Royal Astronomical Society* 224.1 (1987), pp. 13–39.

[346] Wanda DeLandro-Clarke and Gary T Jarvis. "Numerical models of mantle convection with secular cooling". In: *Geophy. J. Int.* 129.1 (1997), pp. 183–193. DOI: 10.1111/j.1365-246X.1997.tb00948.x.

[347] Fábio CG DeMarco, Cláudia R DeAndrade, and Edson L Zaparoli. "The no-slip boundary condition in the stream function-vorticity formulation using the penalty method". In: *International communications in heat and mass transfer* 30.4 (2003), pp. 495–504. DOI: 10.1016/S0735-1933(03)00078-2.

[348] SCR Dennis and JD Hudson. "Methods of solution of the velocity-vorticity formulation of the Navier-Stokes equations". In: *J. Comp. Phys.* 122.2 (1995), pp. 300–306.

[349] S Dequand et al. "Simplified models of flue instruments: Influence of mouth geometry on the sound source". In: *The Journal of the Acoustical Society of America* 113.3 (2003), pp. 1724–1735. DOI: 10.1121/1.1543929.

[350] Y. Deubelbeiss and B.J.P. Kaus. "Comparison of Eulerian and Lagrangian numerical techniques for the Stokes equations in the presence of strongly varying viscosity". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 92–111. DOI: 10.1016/j.pepi.2008.06.023.

[351] Byron DeVries, Joe Iannelli, Christian Trefftz, Kurt A O'Hearn, and Greg Wolffe. "Parallel implementations of FGMRES for solving large, sparse non-symmetric linear systems". In: *Procedia Computer Science* 18 (2013), pp. 491–500.

[352] G. Dhatt and G. Hubert. "A study of penalty elements for incompressible laminar flows". In: *Int. J. Num. Meth. Fluids* 6 (1986), pp. 1–19.

[353] E Di Giuseppe et al. "Characterization of Carbopol® hydrogel rheology for experimental tectonics and geodynamics". In: *Tectonophysics* 642 (2015), pp. 29–45.

[354] Min Ding, Jian Lin, Chen Gu, Qinghua Huang, and Maria T Zuber. "Variations in Martian lithospheric strength based on gravity/topography analysis". In: *J. Geophys. Res.: Planets* 124.11 (2019), pp. 3095–3118. DOI: 10.1029/2019JE005937.

[355] I. Dione, C. Tibirna, and J. Urquiza. "Stokes equations with penalised slip boundary conditions". In: *International Journal of Computational Fluid Dynamics* 27.6-7 (2013), pp. 283–296. DOI: 10.1080/10618562.2013.821114.

[356] Nathaniel A Dixon and William B Durham. "Measurement of activation volume for creep of dry olivine at upper-mantle conditions". In: *J. Geophys. Res.: Solid Earth* 123.10 (2018), pp. 8459–8473. DOI: 10.1029/2018JB015853.

[357] James M Dohm, Shigenori Maruyama, Motoyuki Kido, and Victor R Baker. "A possible anorthositic continent of early Mars and the role of planetary size for the inception of Earth-like life". In: *Geoscience Frontiers* 9.4 (2018), pp. 1085–1098. DOI: 10.1016/j.gsf.2016.12.003.

[358] C.R. Dohrmann and P.B. Bochev. "A stabilized finite element method for the Stokes problem based on polynomial pressure projections". In: *Int. J. Num. Meth. Fluids* 46 (2004), pp. 183–201. DOI: 10.1002/fld.752.

[359] V. Dolejsi and M. Feistauer. *Discontinuous Galerkin Methods. Analysis and Applications to Compressible Flow.* Springer, 2015.

[360] Vít Dolejší. "Analysis and application of the IIPG method to quasilinear nonstationary convection–diffusion problems". In: *Journal of Computational and Applied Mathematics* 222.2 (2008), pp. 251–273.

[361] J Donea, Antonio Huerta, J-Ph Ponthot, and A Rodriguez-Ferran. "Arbitrary Lagrangian-Eulerian Methods, volume 1 of Encyclopedia of Computational Mechanics, chapter 14". In: 3 (2004), pp. 1–25.

[362] J. Donea, S. Giuliani, K. Morgan, and L. Quartapelle. "The significance of chequerboarding in a Galerkin finite element solution of the Navier-Stokes equations". In: *Int. J. Num. Meth. Eng.* 17.5 (1981), pp. 790–795. DOI: 10.1002/nme.1620170511.

[363] Jean Donea and Antonio Huerta. *Finite Element Methods for Flow Problems.* John Wiley & Sons, 2003. ISBN: 978-0-471-49666-3.

[364] J.R. Dormand and P.J. Prince. "A family of embedded Runge-Kutta formulae". In: *Journal of Computational and Applied Mathematics* 6.1 (1980), pp. 19–26.

[365] J.R. Dormand and P.J. Prince. "A reconsideration of some embedded Runge-Kutta formulae". In: *Journal of Computational and Applied Mathematics* 15 (1986), pp. 203–211.

[366] Jim Douglas and Junping Wang. "An absolutely stabilised finite element method for the Stokes problem". In: *Mathematics of Computation* 52.186 (1989), pp. 495–508. DOI: 10.1090/S0025-5718-1989-0958871-X.

[367] Jim Douglas Jr, Juan E Santos, Dongwoo Sheen, and Xiu Ye. "Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 33.4 (1999), pp. 747–770. DOI: 10.1051/m2an:1999161.

[368] M van Driel, Lion Krischer, Simon C Stähler, Kambod Hosseini, and Tarje Nissen-Meyer. "Instaseis: Instant global seismograms based on a broadband waveform database". In: *Solid Earth* 6.2 (2015), pp. 701–717. DOI: 10.5194/se-6-701-2015.

[369] Mélanie Drilleau et al. "Constraints on lateral variations of Martian crustal thickness from seismological and gravity field measurements". In: *Geophys. Res. Lett.* 51.4 (2024), e2023GL105701. DOI: 10.1029/2023GL105701.

[370] D.C. Drucker and W. Prager. "Soil mechanics and plastic analysis or limit design". In: *Quarterly of Applied Mathematics* 10.2 (1952), pp. 157–165. DOI: xxxx.

[371] CP Dubey and VM Tiwari. "Computation of the gravity field and its gradient: Some applications". In: *Computers and Geosciences* 88 (2016), pp. 83–96. DOI: 10.1016/j.cageo.2015.12.007.

[372] J.K. Dukowicz. "Efficient Volume Computation for Three- Dimensional Hexahedral Cells". In: *J. Comp. Phys.* 74 (1988), pp. 493–496. DOI: 10.1016/0021-9991(88)90091-5.

[373] D.A. Dunavant. "High-degree efficient symmetrical Gaussian quadrature rules for the triangle". In: *Int. J. Num. Meth. Eng.* 21 (1985), pp. 1129–1148. DOI: 10.1002/nme.1620210612.

[374] J.A. Dunbar and D.S. Sawyer. "Three-dimensional dynamical model of continental rift propagation and margin plateau formation". In: *J. Geophys. Res.: Solid Earth* 101.B12 (1996), pp. 27, 845–27, 863. DOI: 10.1029/96JB01231.

[375] T. Duretz, D.A. May, T.V. Gerya, and P.J. Tackley. "Discretization errors and free surface stabilisation in the finite difference and marker-in-cell method for applied geodynamics: A numerical study". In: *Geochem. Geophys. Geosyst.* 12 (2011), Q07004. DOI: 10.1029/2011GC003567.

[376] Thibault Duretz, Dave A May, and Philippe Yamato. "A free surface capturing discretization for the staggered grid finite difference scheme". In: *Geophy. J. Int.* 204.3 (2016), pp. 1518–1530. DOI: 10.1093/gji/ggv526.

[377] A. Düster and E. Rank. "The p-version of the finite element method compared to an adaptive h-version for the deformation theory of plasticity". In: *Computer Methods in Applied Mechanics and Engineering* 190 (2011), pp. 1925–1935. DOI: 10.1016/S0045-7825(00)00215-2.

[378] Urmi Dutta, Shamik Sarkar, and Nibir Mandal. "Ballooning versus curling of mantle plumes: views from numerical models". In: *Current Science* 104.7 (2013), pp. 893–903. DOI: xxxx.

[379] S. Dyksterhuis, P. Rey, R.D. Mueller, and L. Moresi. "Effects of initial weakness on rift architecture". In: *Geological Society, London, Special Publications* 282 (2007), pp. 443–455. DOI: 10.1144/SP282.18.

[380] A.M. Dziewonski and D.L. Anderson. "Preliminary reference Earth model". In: *Phys. Earth. Planet. Inter.* 25 (1981), pp. 297–356. DOI: 10.1016/0031-9201(81)90046-7.

[381] J Ebbing, J Bouman, M Fuchs, S Gradmann, and R Haagmans. "Sensitivity of GOCE gravity gradients to crustal thickness and density variations: Case study for the Northeast Atlantic Region". In: *Gravity, Geoid and Height Systems.* Springer, 2014, pp. 291–298. DOI: 10.1007/978-3-319-10837-7_37.

[382] Jörg Ebbing et al. "Advancements in satellite gravity gradient data for crustal studies". In: *The Leading Edge* 32.8 (2013), pp. 900–906. DOI: xxxx.

[383] David L. Egholm. "A new strategy for discrete element numerical models: 1. Theory". In: *J. Geophys. Res.: Solid Earth* 112 (2007), B05203. DOI: 10.1029/2006JB004557.

[384] David L. Egholm, Mike Sandiford, Ole R. Clausen, and Søren B. Nielsen. "A new strategy for discrete element numerical models: 2. Sandbox applications". In: *J. Geophys. Res.: Solid Earth* 112 (2007), B05204. DOI: 10.1029/2006JB004558.

[385] Yuzuru Eguchi. "A new positive-definite regularization of incompressible Navier–Stokes equations discretized with Q1/P0 finite element". In: *International journal for numerical methods in fluids* 41.8 (2003), pp. 881–904. DOI: 10.1002/d.482.

[386] Louis W Ehrlich and Murli M Gupta. "Some difference schemes for the biharmonic equation". In: *SIAM Journal on Numerical Analysis* 12.5 (1975), pp. 773–790. DOI: 10.1137/0712058.

[387] V. Eijkhout. *Introduction to High Performance Scientific Computing.* Creative Commons, 2013.

[388] Stanley C Eisenstat, Howard C Elman, and Martin H Schultz. "Variational iterative methods for nonsymmetric systems of linear equations". In: *SIAM Journal on Numerical Analysis* 20.2 (1983), pp. 345–357.

[389] Dirk Elbeshausen and Jay Melosh. "A nonlinear and time-dependent visco-elasto-plastic rheology model for studying shock-physics phenomena". In: *arXiv preprint arXiv:1805.06453* (2018).

[390] Y. Elesin, T. Gerya, I.M. Artemieva, and H. Thybo. "Samovar: a thermomechanical code for modeling of geodynamic processes in the lithospshere – application to basin evolution". In: *Arabian Journal of Geosciences* 3 (2010), pp. 477–497. DOI: 10.1007/s12517-010-0215-1.

[391] S. Ellis, P. Fullsack, and C. Beaumont. "Oblique convergence of the crust driven by basal forcing: implications for length-scales of deformation and strain partitioning in orogens". In: *Geophy. J. Int.* 120 (1995), pp. 24–44. DOI: 10.1111/j.1365-246X.1995.tb05909.x.

[392] Susan Ellis, Guido Schreurs, and Marion Panien. "Comparisons between analogue and numerical models of thrust wedge development". In: *Journal of Structural Geology* 26.9 (2004), pp. 1659–1675. DOI: 10.1016/j.jsg.2004.02.012.

[393] Kirk Ellsworth, Gerald Schubert, and Charles G Sammis. "Viscosity profile of the lower mantle". In: *Geophy. J. Int.* 83.1 (1985), pp. 199–213. DOI: 10.1111/j.1365-246X.1985.tb05163.x.

[394] H. Elman, D. Silvester, and A. Wathen. *Finite Elements and Fast Iterative Solvers*. Oxford Science Publications, 2014. ISBN: 978-0-19-967879-2.

[395] M.S. Engelman, R.L. Sani, and P.M. Gresho. "The implementation of normal and/or tangential boundary conditions in finite element codes for incompressible fluid flow". In: *Int. J. Num. Meth. Fluids* 2 (1982), pp. 225–238. DOI: 10.1002/fld.1650020302.

[396] MS Engelman, Gilbert Strang, and K-J Bathe. "The application of quasi-Newton methods in fluid mechanics". In: *Int. J. Num. Meth. Eng.* 17.5 (1981), pp. 707–718.

[397] P. England. "Some numerical investigations of large scale continental deformation". In: *Mountain Building Processes*. Academic Press, 1982, pp. 129–189. DOI: xxxx.

[398] P. England and G. Houseman. "A dynamical model of lithosphere extension and sedimentary basin formation". In: *J. Geophys. Res.: Solid Earth* 91.B3 (1986), pp. 3664–3676. DOI: 10.1029/JB091iB01p00719.

[399] Philip England and Dan McKenzie. "A thin viscous sheet model for continental deformation". In: *Geophy. J. Int.* 70.2 (1982), pp. 295–321. DOI: 10.1111/j.1365-246X.1982.tb04969.x.

[400] Philip England and Peter Molnar. "Active deformation of Asia: From kinematics to dynamics". In: *Science* 278.5338 (1997), pp. 647–650. DOI: 10.1126/science.278.5338.647.

[401] Philip England and Catherine Wilkins. "A simple analytical approximation to the temperature structure in subduction zones". In: *Geophy. J. Int.* 159.3 (2004), pp. 1138–1154. DOI: 10.1111/j.1365-246X.2004.02419.x.

[402] I. Ergatoudis, B.M. Irons, and O.C. Zienkiewicz. "Curved, isoparametric, "quadrilateral" elements for finite element analysis". In: *International journal of solids and structures* 4.1 (1968), pp. 31–42. DOI: 10.1016/0020-7683(68)90031-0.

[403] VV Ermakov and Nikolai Nikolaevich Kalitkin. "The optimal step and regularization for Newton's method". In: *USSR Computational Mathematics and Mathematical Physics* 21.2 (1981), pp. 235–242. DOI: 10.1016/0041-5553(81)90022-7.

[404] Alexandre Ern and Jean-Luc Guermond. *Finite elements I: Approximation and interpolation*. Vol. 72. Springer Nature, 2021.

[405] E. Erturk. "Discussions on Driven Cavity Flow". In: *Int. J. Num. Meth. Fluids* 60 (2009), pp. 275–294.

[406] Pep Espanol and Cedric Thieulot. "Microscopic derivation of hydrodynamic equations for phase-separating fluid mixtures". In: *The Journal of Chemical Physics* 118.20 (2003), pp. 9109–9127. DOI: 10.1063/1.1568333.

[407] Grant Thomas Euen, Shangxin Liu, Rene Gassmöller, Timo Heister, and Scott David King. "A comparison of 3-D spherical shell thermal convection results at low to moderate Rayleigh number using ASPECT (version 2.2. 0) and CitcomS (version 3.3. 1)". In: *Geosci. Model. Dev.* 2022 (2023), pp. 1–34. DOI: 10.5194/gmd-16-3221-2023.

[408] Brian Evans and Christopher Goetze. "The temperature variation of hardness of olivine and its implication for polycrystalline yield stress". In: *J. Geophys. Res.: Solid Earth* 84.B10 (1979), pp. 5505–5524. DOI: 10.1029/JB084iB10p05505.

[409] M. Faccenda, T.V. Gerya, N.S. Mancktelow, and L. Moresi. "Fluid flow during slab unbending and dehydration: Implications for intermediate-depth seismicity, slab weakening and deep water recycling". In: *Geochem. Geophys. Geosyst.* 13.1 (2012). DOI: 10.1029/2011GC003860.

[410] Cinzia G Farnetani and Henri Samuel. "Lagrangian structures and stirring in the Earth's mantle". In: *Earth Planet. Sci. Lett.* 206.3-4 (2003), pp. 335–348. DOI: 10.1016/S0012-821X(02)01085-3.

[411] Patrick E Farrell, Lawrence Mitchell, L Ridgway Scott, and Florian Wechsung. "A Reynolds-robust preconditioner for the Scott-Vogelius discretization of the stationary incompressible Navier-Stokes equations". In: *The SMAI journal of computational mathematics* 7 (2021), pp. 75–96. DOI: 10.5802/smai-jcm.72.

[412] R.J. Farrington, L.-N. Moresi, and F.A. Capitanio. "The role of viscoelasticity in subducting plates". In: *Geochem. Geophys. Geosyst.* 15 (2014), pp. 4291–4304. DOI: 10.1002/2014GC005507.

[413] U.H. Faul, J.D. Fitz Gerald, R. J.M. Farlai, R. Ahlefeldt, and I. Jackson. "Dislocation creep of fine-grained olivine". In: *J. Geophys. Res.: Solid Earth* 116.B01203, (2011). DOI: 10.1029/2009JB007174.

[414] E. Fehlberg. "Some old and new Runge-Kutta formulas with stepsize control and their error coefficients". In: *Computing* 34 (1985), pp. 265–270. DOI: 10.1007/BF02253322.

[415] Marc Fehling and Wolfgang Bangerth. "Algorithms for parallel generic hp-adaptive finite element software". In: *ACM Transactions on Mathematical Software* 49.3 (2023), pp. 1–26. DOI: 10.1145/3603372.

[416] JA Fernández-Merodo, JC García-Davalillo, G Herrera, P Mira, and M Pastor. "2D viscoplastic finite element modelling of slow landslides: the Portalet case study (Spain)". In: *Landslides* 11.1 (2014), pp. 29–42. DOI: 10.1007/s10346-012-0370-4.

[417] Sylvaine Ferrachat and Yanick Ricard. "Mixing properties in the Earth's mantle: Effects of the viscosity stratification and of oceanic crust segregation". In: *Geochem. Geophys. Geosyst.* 2.4 (2001). DOI: 10.1029/2000GC000092.

[418] Sylvaine Ferrachat and Yanick Ricard. "Regular vs. chaotic mantle mixing". In: *Earth Planet. Sci. Lett.* 155.1-2 (1998), pp. 75–86. DOI: 10.1016/S0012-821X(97)00200-8.

[419] Manuel E Fiadeiro and George Veronis. "On weighted-mean schemes for the finite-difference approximation to the advection-diffusion equation". In: *Tellus* 29.6 (1977), pp. 512–522. DOI: 10.3402/tellusa.v29i6.11385.

[420] David A Field. "Qualitative measures for initial meshes". In: *Int. J. Num. Meth. Eng.* 47.4 (2000), pp. 887–906. DOI: 10.1002/(SICI)1097-0207(20000210)47:4<887::AID-NME804>3.0.CO;2-H.

[421] Mark P Fischer, Michael R Gross, Terry Engelder, and Roy J Greenfield. "Finite-element analysis of the stress distribution around a pressurized crack in a layered elastic medium: implications for the spacing of fluid-driven joints in bedded sedimentary rock". In: *Tectonophysics* 247.1-4 (1995), pp. 49–64. DOI: 10.1016/0040-1951(94)00200-S.

[422] Nicolas Flament. "Present-day dynamic topography and lower-mantle structure from palaeogeographically constrained mantle flow models". In: *Geophy. J. Int.* 216.3 (2019), pp. 2158–2182. DOI: 10.1093/gji/ggy526.

[423] Luce Fleitout and David A Yuen. "Secondary convection and the growth of the oceanic lithosphere". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 181–212. DOI: 10.1016/0031-9201(84)90046-3.

[424] Raymond C Fletcher. "Three-dimensional folding of an embedded viscous layer in pure shear". In: *Journal of Structural Geology* 13.1 (1991), pp. 87–96.

[425] MGG Foreman and AF Bennett. "On no-slip boundary conditions for the incompressible Navier-Stokes equations". In: *Dynamics of atmospheres and oceans* 12.1 (1988), pp. 47–70. DOI: 10.1016/0377-0265(88)90014-0.

[426] A. Fortin, M. Jardak, J.J. Gervais, and R. Pierre. "Old and New Results on the Two-Dimensional Poiseuille Flow". In: *J. Comp. Phys.* 115.2 (1994), pp. 455–469. DOI: 10.1006/jcph.1994.1210.

[427] M. Fortin. "Old and new finite elements for incompressible flows". In: *Int. J. Num. Meth. Fluids* 1 (1981), pp. 347–364. DOI: 10.1002/fld.1650010406.

[428] M. Fortin and S. Boivin. "Iterative stabilisation of the bilinear velocity-constant pressure element". In: *Int. J. Num. Meth. Fluids* 10 (1990), pp. 125–140. DOI: 10.1002/fld.1650100202.

[429] M. Fortin and A. Fortin. "Experiments with several elements for viscous incompressible flows". In: *Int. J. Num. Meth. Fluids* 5 (1985), pp. 911–928. DOI: 10.1002/fld.1650051005.

[430] M. Fortin and M. Soulie. "A non-conforming piecewise quadratic finite element on triangles". In: *Int. J. Num. Meth. Eng.* 19 (1983), pp. 505–520.

[431] Michel Fortin and F Thomasset. "Mixed finite-element methods for incompressible flow problems". In: *J. Comp. Phys.* 31.1 (1979), pp. 113–145.

[432] L. Fourel, S. Goes, and G. Morra. "The role of elasticity in slab bending". In: *Geochem. Geophys. Geosyst.* 15 (2014), pp. 4507–4525. DOI: 10.1002/2014GC005535.

[433] Marc Fournier, Laurent Jolivet, Philippe Davy, and Jean-Charles Thomas. "Backarc extension and collision: an experimental approach to the tectonics of Asia". In: *Geophy. J. Int.* 157.2 (2004), pp. 871–889.

[434] Christine Mary Rutherford Fowler. *The solid earth: an introduction to global geophysics; second edition.* Cambridge University Press, 2005.

[435] L.P. Franca and T.J.R. Hughes. "Convergence analyses of Galerkin least-squares methods for symmetric advective-diffusive forms of the Stokes and incompressible Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 105 (1993), pp. 285–298. DOI: 10.1016/0045-7825(93)90126-I.

[436] L.P. Franca and S.P. Oliveira. "Pressure bubbles stabilization features in the Stokes problem". In: *Computer Methods in Applied Mechanics and Engineering* 192 (2003), pp. 1929–1937.

[437] L.P. Franca, S.P. Oliveira, and M. Sarkis. "Continuous Q1-Q1 Stokes elements stabilised with non-conforming null edge average velocity functions". In: *Mathematical Models and Methods in Applied Sciences* 17 (03 2007), pp. 439–459. DOI: 10.1142/S021820250700198X.

[438] Leopoldo P Franca, Sergio L Frey, and Thomas JR Hughes. "Stabilized finite element methods: I. Application to the advective-diffusive model". In: *Computer Methods in Applied Mechanics and Engineering* 95.2 (1992), pp. 253–276. DOI: 10.1016/0045-7825(92)90143-8.

[439] Leopoldo P Franca, G Hauke, and A Masud. "Stabilized finite element methods". In: *FINITE ELEMENT METHODS: 1970's AND BEYOND.* International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 2004.

[440] M. Fraters, C. Thieulot, A. van den Berg, and W. Spakman. "The Geodynamic World Builder: a solution for complex initial conditions in numerical modelling". In: *Solid Earth* 10 (2019), pp. 1785–1807. DOI: 10.5194/se-10-1785-2019.

[441] M.R.T. Fraters, W. Bangerth, C. Thieulot, A.C. Glerum, and W. Spakman. "Efficient and Practical Newton Solvers for Nonlinear Stokes Systems in Geodynamic Problems". In: *Geophy. J. Int.* 218 (2019), pp. 873–894. DOI: 10.1093/gji/ggz183.

[442] S. Frederiksen and J. Braun. "Numerical modelling of strain localisation during extension of the continental lithosphere". In: *Earth Planet. Sci. Lett.* 188 (2001), pp. 241–251. DOI: 10.1016/S0012-821X(01)00323-5.

[443] M. Frehner. "3D fold growth rates". In: *Terra Nova* 26 (2014), pp. 417–424. DOI: 10.1111/ter.12116.

[444] Alfred M. Freudenthal and Hilda Geiringer. "The mathematical theory of the inelastic continuum". In: *Handbuch der Physik, Encyclopedia of Physics.* Vol. VI. Springer-Verlag, 1958.

[445] P.J. Frey and P.-L. George. *Mesh generation.* Hermes Science, 2000.

[446] C. Froidevaux. "Energy dissipation and geometric structure at spreading plate boundaries". In: *Earth Planet. Sci. Lett.* 20 (1973), pp. 419–424. DOI: 10.1016/0012-821X(73)90020-4.

[447] L. Fuchs and H. Schmeling. "A new numerical method to calculate inhomogeneous and time-dependent large deformation of two-dimensional geodynamic flows with application to diapirism". In: *Geophy. J. Int.* 194.2 (2013), pp. 623–639. DOI: 10.1093/gji/ggt142.

[448] Hiromi Fujimoto and Yoshibumi Tomoda. "Lithospheric thickness anomaly near the trench and possible driving force of subduction". In: *Tectonophysics* 112.1-4 (1985), pp. 103–110. DOI: 10.1016/0040-1951(85)90174-X.

[449] T. Fukuchi. "Numerical calculation of fully-developed laminar flows in arbitrary cross-sections using finite difference method". In: *AIP Advances* 1 (2011), p. 042109. DOI: 10.1063/1.3652881.

[450] Javier Fullea, Juan Rodríguez-González, María Charco, Zdenek Martinec, A Negredo, and Antonio Villaseñor. "Perturbing effects of sub-lithospheric mass anomalies in GOCE gravity gradient and other gravity data modelling: Application to the Atlantic-Mediterranean transition zone". In: *International Journal of Applied Earth Observation and Geoinformation* 35 (2015), pp. 54–69. DOI: 10.1016/j.jag.2014.02.003.

[451] CW Fuller, SD Willett, D Fisher, and CY Lu. "A thermomechanical wedge model of Taiwan constrained by fission-track thermochronometry". In: *Tectonophysics* 425.1-4 (2006), pp. 1–24. DOI: 10.1016/j.tecto.2006.05.018.

[452] P. Fullsack. "An arbitrary Lagrangian-Eulerian formulation for creeping flows and its application in tectonic models". In: *Geophy. J. Int.* 120 (1995), pp. 1–23. DOI: 10.1111/j.1365-246X.1995.tb05908.x.

[453] Jean Furstoss. "Approche numérique de l'évolution microstructurale des péridotites". PhD thesis. 2020.

[454] Mikito Furuchi. "Numerical modeling of three dimensional self-gravitating Stokes flow problem with free surface". In: *Procedia Computer Science* 4 (2011), pp. 1506–1515. DOI: 10.1016/j.procs.2011.04.163.

[455] M. Furuichi, M. Kameyama, and A. Kageyama. "Three-dimensional Eulerian method for large deformation of viscoelastic fluid: Toward plate-mantle simulation". In: *J. Comp. Phys.* 227 (2008), pp. 4977–4997. DOI: 10.1016/j.jcp.2008.01.052.

[456] M. Furuichi and D. Nishiura. "Robust coupled fluid-particle simulation scheme in Stokes-flow regime: Toward the geodynamic simulation including granular media". In: *Geochem. Geophys. Geosyst.* 15 (2014), pp. 2865–2882.

[457] Mikito Furuichi, Dave A May, and Paul J Tackley. "Development of a Stokes flow solver robust to large viscosity jumps using a Schur complement approach with mixed precision arithmetic". In: *J. Comp. Phys.* 230.24 (2011), pp. 8835–8851. DOI: `10.1016/j.jcp.2011.09.007`.

[458] Yoshitsugu Furukawa. "Depth of the decoupling plate interface and thermal structure under arcs". In: *J. Geophys. Res.: Solid Earth* 98.B11 (1993), pp. 20005–20013.

[459] Carl W Gable, Richard J O'connell, and Bryan J Travis. "Convection in three dimensions with surface plates: Generation of toroidal flow". In: *J. Geophys. Res.: Solid Earth* 96.B5 (1991), pp. 8391–8405. DOI: `10.1029/90JB02743`.

[460] Sashikumaar Ganesan, Gunar Matthies, and Lutz Tobiska. "Local projection stabilization of equal order interpolation applied to the Stokes problem". In: *Mathematics of Computation* 77.264 (2008), pp. 2039–2060. DOI: `10.1090/S0025-5718-08-02130-3`.

[461] F. Garel, S. Goes, D.R. Davies, J.H. Davies, S.C. Kramer, and C.R. Wilson. "Interaction of subducted slabs with the mantle transition-zone: A regime diagram from 2-D thermo-mechanical models with a mobile trench and an overriding plate". In: *Geochem. Geophys. Geosyst.* 15.1739–1765 (2014). DOI: `10.1002/2014GC005257`.

[462] Rao Garimella. "Conformal refinement of unstructured quadrilateral meshes". In: *Proceedings of the 18th International Meshing Roundtable*. Springer, 2009, pp. 31–44.

[463] D.K. Gartling. *Nachos - A finite element computer program for incompressible flow problems.* Tech. rep. Sand77-1333. Sandia Laboratories, 1978.

[464] R. Gassmöller, H. Lokavarapu, E. M. Heien, E. G. Puckett, and W. Bangerth. "Flexible and scalable particle-in-cell methods with adaptive mesh refinement for geodynamic computations". In: *Geochem. Geophys. Geosyst.* 19.9 (2018), pp. 3596–3604. DOI: `10.1029/2018GC007508`.

[465] Rene Gassmöller, Juliane Dannberg, Wolfgang Bangerth, Timo Heister, and Robert Myhill. "On formulations of compressible mantle convection". In: *Geophy. J. Int.* 221.2 (2020), pp. 1264–1280. DOI: `10.1093/gji/ggaa078`.

[466] Rene Gassmöller, Juliane Dannberg, Wolfgang Bangerth, Elbridge Gerry Puckett, and Cedric Thieulot. "Benchmarking the accuracy of higher order particle methods in geodynamic models of transient flow". In: *Geosci. Model. Dev.* 17 (2024), pp. 4115–4134. DOI: `10.5194/gmd-17-4115-2024`.

[467] Rene Gassmöller, Harsha Lokavarapu, Wolfgang Bangerth, and Elbridge Gerry Puckett. "Evaluating the accuracy of hybrid finite element/particle-in-cell methods for modelling incompressible Stokes flow". In: *Geophy. J. Int.* 219.3 (2019), pp. 1915–1938. DOI: `10.1093/gji/ggz405`.

[468] L. Gastaldo. "Methodes de correction de pression pour les ecoulements compressibles: Application aux equations de Navier-Stokes barotropes et au modele de derive". PhD thesis. Universite de Provence, 2007.

[469] Gabriel N Gatica. "A simple introduction to the mixed finite element method". In: *Theory and Applications. Springer Briefs in Mathematics. Springer, London* (2014).

[470] Andrey A Gavrilov, Konstantin A Finnikov, and Evgeny V Podryabinkin. "Modeling of steady Herschel–Bulkley fluid flow over a sphere". In: *Journal of Engineering Thermophysics* 26.2 (2017), pp. 197–215. DOI: `10.1134/S1810232817020060`.

[471] T. Geenen et al. "Scalable robust solvers for unstructured FE geodynamic modeling applications: Solving the Stokes equation for models with large localized viscosity contrasts". In: *Geochem. Geophys. Geosyst.* 10.9 (2009). DOI: `10.1029/2009GC002526`.

[472]   P.-L. George and H. Borouchaki. *Delaunay Triangulation and Meshing.* Hermes, 1998.

[473]   I. Georgiev, J. Kraus, and S. Margenov. "Multilevel algorithms for Rannacher-Turek finite element approximations of 3D elliptic problems". In: *Computing* 82 (2008), pp. 217–239. DOI: 10.1007/s00607-008-0008-5.

[474]   M. Gerault, T.W. Becker, B.J.P. Kaus, C. Faccenna, L. Moresi, and L. Husson. "The role of slabs and oceanic plate geometry in the net rotation of the lithosphere, trench motions, and slab return flow". In: *Geochem. Geophys. Geosyst.* 13.4 (2012), Q04001. DOI: 10.1029/2011GC003934.

[475]   M. Gerbault. "Pressure conditions for shear and tensile failure around a circular magma chamber; insight from elasto-plastic modelling". In: *Geological Society, London, Special Publications* 367 (2012), pp. 111–130.

[476]   M. Gerbault, F. Cappa, and R. Hassani. "Elasto-plastic and hydromechanical models of failure around an infinitely long magma chamber". In: *Geochem. Geophys. Geosyst.* 13.3 (2012). DOI: 10.1029/2011GC003917.

[477]   M. Gerbault, A.N.B. Poliakov, and M. Daignieres. "Prediction of faulting from the theories of elasticity and plasticity: what are the limits?" In: *Journal of Structural Geology* 20 (1998), pp. 301–320.

[478]   T. Gerya and D.A. Yuen. "Robust characteristics method for modelling multiphase visco-elasto-plastic thermo-mechanical problems". In: *Phys. Earth. Planet. Inter.* 163 (2007), pp. 83–105. DOI: 10.1016/j.pepi.2007.04.015.

[479]   T.V. Gerya and J.-P. Burg. "Intrusion of ultramafic magmatic bodies into the continental crust: Numerical simulation". In: *Phys. Earth. Planet. Inter.* 160 (2007), pp. 124–142. DOI: 10.1016/j.pepi.2006.10.004.

[480]   T.V. Gerya, D.A. May, and T. Duretz. "An adaptive staggered grid finite difference method for modeling geodynamic Stokes flows with strongly variable viscosity". In: *Geochem. Geophys. Geosyst.* 14.4 (2013), pp. 1200–1225. DOI: 10.1002/ggge.20078.

[481]   T.V. Gerya and D.A. Yuen. "Characteristics-based marker-in-cell method with conservative finite-differences schemes for modeling geological flows with strongly variable transport properties". In: *Phys. Earth. Planet. Inter.* 140 (2003), pp. 293–318. DOI: 10.1016/j.pepi.2003.09.006.

[482]   Taras Gerya. "Geodynamics of the early Earth: Quest for the missing paradigm". In: *Geology* 47.10 (2019), pp. 1006–1007. DOI: 10.1130/focus102019.1.

[483]   Taras Gerya. *Numerical Geodynamic Modelling.* Cambridge University Press, 2010.

[484]   Taras Gerya. *Numerical Geodynamic Modelling - 2nd edition.* Cambridge University Press, 2019. ISBN: 978-1-107-14314-2.

[485]   Christophe Geuzaine and Jean-François Remacle. "Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities". In: *International journal for numerical methods in engineering* 79.11 (2009), pp. 1309–1331. DOI: 10.1002/nme.2579.

[486]   Siavash Ghelichkhan and Hans-Peter Bunge. "The compressible adjoint equations in geodynamics: derivation and numerical assessment". In: *GEM-International Journal on Geomathematics* 7.1 (2016), pp. 1–30. DOI: 10.1007/s13137-016-0080-5.

[487]   U. Ghia, K.N. Ghia, and C.T. Shin. "High-Re Solutions for incompressible flow using the Navier-Stokes equations and a multigrid method". In: *J. Comp. Phys.* 48 (1982), pp. 387–411.

[488] Norman E Gibbs. "Algorithm 509: A hybrid profile reduction algorithm [F1]". In: *ACM Transactions on Mathematical Software (TOMS)* 2.4 (1976), pp. 378–387.

[489] Norman E Gibbs, William G Poole Jr, and Paul K Stockmeyer. "A comparison of several bandwidth and profile reduction algorithms". In: *ACM Transactions on Mathematical Software (TOMS)* 2.4 (1976), pp. 322–330. DOI: `xxxx`.

[490] Frederic Gibou, Ronald Fedkiw, and Stanley Osher. "A review of level-set methods and some recent applications". In: *J. Comp. Phys.* 353 (2018), pp. 82–109. DOI: `10.1016/j.jcp.2017.10.006`.

[491] JA Gil and Maria José Jurado. "Geological interpretation and numerical modelling of salt movement in the Barbastro–Balaguer anticline, southern Pyrenees". In: *Tectonophysics* 293.3-4 (1998), pp. 141–155.

[492] Freeman Gilbert and George E Backus. "Propagator matrices in elastic wave and vibration problems". In: *Geophysics* 31.2 (1966), pp. 326–332. DOI: `10.1190/1.1439771`.

[493] François Glaisner and Tayfun E Tezduyar. *Finite element techniques for the Navier-Stokes equations in the primitive variable formulation and the vorticity stream-function formulation.* Tech. rep. 1987.

[494] G.C. Gleason and J. Tullis. "A flow law for dislocation creep of quartz aggregates determined with the molten salt cell". In: *Tectonophysics* 247 (1995), pp. 1–23.

[495] John W Glen. "The creep of polycrystalline ice". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 228.1175 (1955), pp. 519–538. DOI: `10.1098/rspa.1955.0066`.

[496] A. Glerum, C. Thieulot, M. Fraters, C. Blom, and W. Spakman. "Nonlinear viscoplasticity in ASPECT: benchmarking and applications to subduction". In: *Solid Earth* 9.2 (2018), pp. 267–294. DOI: `10.5194/se-9-267-2018`.

[497] P Glišović, AM Forte, and Robert Moucha. "Time-dependent convection models of mantle thermal structure constrained by seismic tomography and geodynamics: implications for mantle plume dynamics and CMB heat flux". In: *Geophy. J. Int.* 190.2 (2012), pp. 785–815. DOI: `10.1111/j.1365-246X.2012.05549.x`.

[498] Björn Gmeiner, Markus Huber, Lorenz John, Ulrich Rüde, and Barbara Wohlmuth. "A quantitative performance study for Stokes solvers at the extreme scale". In: *Journal of Computational Science* 17 (2016), pp. 509–521. DOI: `10.1016/j.jocs.2016.06.006`.

[499] C. Goetze and B. Evans. "Stress and temperature in the bending lithosphere as constrained by experimental rock mechanics". In: *Geophy. J. Int.* 59.3 (1979), pp. 463–478.

[500] DL Goldsby and DL Kohlstedt. "Superplastic deformation of ice: Experimental observations". In: *J. Geophys. Res.: Solid Earth* 106.B6 (2001), pp. 11017–11030. DOI: `10.1029/2000JB900336`.

[501] G. Goltz and M. Böse. "Configurational entropy of critical earthquake populations". In: *Geophys. Res. Lett.* 29.20 (2002). DOI: `10.1029/2002GL015540`.

[502] G.H. Golub and C.F. van Loan. *Matrix Computations, 4th edition.* John Hopkins University Press, 2013.

[503] H Gossman. "Slope modelling with changing boundary conditions-effects of climate and lithology". In: *Z. Geomoph. NF* 25 (1976), pp. 72–88.

[504] K-D Gottschaldt, U Walzer, RF Hendel, David Robert Stegman, JR Baumgardner, and H-B Mühlhaus. "Stirring in 3-d spherical models of convection in the Earth's mantle". In: *Philosophical Magazine* 86.21-22 (2006), pp. 3175–3204. DOI: `10.1080/14786430500197991`.

[505] Klaus-D Gottschaldt, Uwe Walzer, Dave R Stegman, John R Baumgardner, and Hans B Mühlhaus. "Mantle Dynamics–A Case Study". In: *Advances in Geocomputing*. 2009, pp. 139–181. DOI: 10.1007/978-3-540-85879-9_5.

[506] S. Gourvenec. "Bearing capacity under combined loading". In: *9th Australia New Zealand Conference on Geomechanics, Auckland, New Zealand, 8-11 february 2004*. 2004.

[507] S. Gourvenec, M. Randolph, and O. Kingsnorth. "Undrained bearing capacity of square and rectangular footings". In: *International Journal of Geomechanics* 6 (2006), pp. 147–157.

[508] R Govers and MJR Wortel. "Initiation of asymmetric extension in continental lithosphere". In: *Tectonophysics* 223.1-2 (1993), pp. 75–96. DOI: 10.1016/0040-1951(93)90159-H.

[509] R. Govers and M.J.R. Wortel. "Lithosphere tearing at STEP faults: Response to edges of subduction zones ". In: *Earth Planet. Sci. Lett.* 236 (2005), pp. 505–523.

[510] J. Grandy. *Efficient Computation of Volume of Hexahedral Cells*. Tech. rep. UCRL-ID-128886. Lawrence Livermore National Laboratory, 1997.

[511] R. Gray and R.N. Pysklywec. "Geodynamic models of mature continental collision: Evolution of an orogen from lithospheric subduction to continental retreat/delamination". In: *J. Geophys. Res.: Solid Earth* 117.B03408 (2012). DOI: 10.1029/2011JB008692.

[512] R. Gray and R.N. Pysklywec. "Influence of viscosity pressure dependence on deep lithospheric tectonics during continental collision". In: *J. Geophys. Res.: Solid Earth* 118 (2013). DOI: 10.1002/jgrb.50220.

[513] Harry W Green. "Shearing instabilities accompanying high-pressure phase transformations and the mechanics of deep earthquakes". In: *Proceedings of the National Academy of Sciences* 104.22 (2007), pp. 9133–9138. DOI: 10.1073pnas.0608045104.

[514] P.M. Gresho, S.T. Chan, M.A. Christon, and A.C. Hindmarsch. "A little more on stabilised Q_1Q_1 for transient viscous incompressible flow". In: *Int. J. Num. Meth. Fluids* 21 (1995), pp. 837–856. DOI: 10.1002/fld.1650211005.

[515] P.M. Gresho and R.L. Lee. "Don't suppress the wiggles - They're telling you something!" In: *Computers and Fluids* 9 (1981), pp. 223–253.

[516] P.M. Gresho, R.L. Lee, R.L. Sani, M.K. Maslanik, and B.E. Eaton. "The consistent Galerkin FEM for computing derived boundary quantities in thermal and/or fluid problems". In: *Int. J. Num. Meth. Fluids* 7 (1987), pp. 371–394. DOI: 10.1002/fld.1650070406.

[517] P.M. Gresho and R.L. Sani. *Incompressible flow and the Finite Element Method, vol II - Isothermal Laminar Flow*. John Wiley and Sons, Ltd, 2000. ISBN: 978-0471492504.

[518] P.M. Gresho and S.B. Sutton. "Application of the FIDAP code to the 8:1 thermal cavity problem". In: *Int. J. Num. Meth. Fluids* 40 (2002), pp. 1083–1092. DOI: 10.1002/d.394.

[519] Philip M Gresho. "Some current CFD issues relevant to the incompressible Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 87.2-3 (1991), pp. 201–252. DOI: 10.1016/0045-7825(91)90006-R.

[520] PM Gresho and RL Lee. "Partial vindication of the bilinear velocity, piecewise constant pressure element". In: *J. Comp. Phys.* 60.1 (1985), pp. 161–164. DOI: 10.1016/0021-9991(85)90023-3.

[521] Ralf Greve. "Application of a polythermal three-dimensional ice sheet model to the Greenland ice sheet: response to steady-state and transient climate scenarios". In: *Journal of Climate* 10.5 (1997), pp. 901–918. DOI: 10.1175/1520-0442(1997)010<0901:AOAPTD>2.0.CO;2.

[522] Ralf Greve and Heinz Blatter. *Dynamics of ice sheets and glaciers.* Springer Science & Business Media, 2009.

[523] M. Griebel, T. Dornseifer, and T. Neunhoeffer. *Numerical simulation in Fluid Dynamics.* SIAM, 1997.

[524] D. Griffiths and D. Silvester. *Unstable modes of the Q1-P0 element.* Tech. rep. 257. University of MAnchester/UMIST, 1994.

[525] D.F. Griffiths. "Finite Elements for Incompressible Flow". In: *Math. Meth. in the Appl. Sci.* 1 (1979), pp. 16–31.

[526] Piotr P Grinevich and Maxim A Olshanskii. "An iterative method for the Stokes-type problem with variable viscosity". In: *SIAM Journal on Scientific Computing* 31.5 (2009), pp. 3959–3978. DOI: 10.1137/08744803.

[527] Thomas Grombein, Kurt Seitz, and Bernhard Heck. "Optimized formulas for the gravitational field of a tesseroid". In: *Journal of Geodesy* 87.7 (2013), pp. 645–660. DOI: 10.1007/s00190-013-0636-1.

[528] L. Gross, L. Bourgouin, A. Hale, and H.-B. Mühlhaus. "Interface modeling in incompressible media using level sets in Escript". In: *Phys. Earth. Planet. Inter.* 163 (2007), pp. 23–34.

[529] Ólafur Gudmundsson and Malcolm Sambridge. "A regionalized upper mantle (RUM) seismic model". In: *J. Geophys. Res.: Solid Earth* 103.B4 (1998), pp. 7121–7136.

[530] J.-L. Guermond, R. Pasquetti, and Bojan Popov. "Entropy viscosity method for nonlinear conservation laws". In: *J. Comp. Phys.* (2011). DOI: 10.1016/j.jcp.2010.11.043.

[531] Jean Luc Guermond, Richard Pasquetti, and Bojan Popov. "Entropy viscosity for conservation equations". In: *V European Conference on Computational Fluid Dynamics (Eccomas CFD 2010).* 2010.

[532] Jean-Luc Guermond and Richard Pasquetti. "Entropy viscosity method for high-order approximations of conservation laws". In: *Spectral and high order methods for partial differential equations.* 2011, pp. 411–418.

[533] JM Guerrero, Julian P Lowman, and Paul J Tackley. "Spurious transitions in convective regime due to viscosity clipping: ramifications for modeling planetary secular cooling". In: *Geochem. Geophys. Geosyst.* 20.7 (2019), pp. 3450–3468. DOI: 10.1029/2019GC008385.

[534] F. Gueydan, C. Morency, and J.-P. Brun. "Continental rifting as a function of lithosphere mantle strength". In: *Tectonophysics* 460 (2008), pp. 83–93. DOI: 10.1016/j.tecto.2008.08.012.

[535] Alex Guinard, Agnes Fienga, Anthony Memin, and Clement Ganino. "Coupled geophysical and thermal constraints link between Mars basal molten layer and the planet viscosity profile". In: *arXiv preprint arXiv:2403.06763* (2024).

[536] G Guj and F Stella. "A vorticity-velocity method for the numerical solution of 3D incompressible flows". In: *J. Comp. Phys.* 106.2 (1993), pp. 286–298. DOI: 10.1016/S0021-9991(83)71108-3.

[537] M. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice and Algorithms.* Academic, Boston, 1989.

[538] Max D Gunzburger and Janet S Peterson. "On finite element approximations of the streamfunction-vorticity and velocity-vorticity equations". In: *Int. J. Num. Meth. Fluids* 8.10 (1988), pp. 1229–1240. DOI: 10.1002/fld.1650081010.

[539] Peng Guo, Leihua Yao, and Desheng Ren. "Simulation of three-dimensional tectonic stress fields and quantitative prediction of tectonic fracture within the Damintun Depression, Liaohe Basin, northeast China". In: *Journal of Structural Geology* 86 (2016), pp. 211–223. DOI: 10.1016/j.jsg.2016.03.007.

[540] Anshul Gupta, George Karypis, and Vipin Kumar. "Highly Scalable Parallel Algorithms for Sparse Matrix Factorization". In: *IEEE Transactions on Parallel and Distributed Systems* 8.5 (1997), pp. 502–520.

[541] Anshul Gupta, Seid Koric, and Thomas George. "Sparse Matrix Factorization on Massively Parallel Computers". In: *SC09 (International Conference for High Performance Computing, Networking, Storage and Analysis)*. 2009.

[542] P.S. Gupta and A.S. Gupta. "Squeezing flow between parallel plates". In: *Wear* 45.2 (1977), pp. 177–185. DOI: 10.1016/0043-1648(77)90072-2.

[543] Michael Gurnis. "Stirring and mixing in the mantle by plate-scale flow: Large persistent blobs and long tendrils coexist". In: *Geophys. Res. Lett.* 13.13 (1986), pp. 1474–1477. DOI: 10.1029/GL013i013p01474.

[544] Michael Gurnis and Geoffrey F Davies. "Mixing in numerical models of mantle convection incorporating plate kinematics". In: *J. Geophys. Res.: Solid Earth* 91.B6 (1986), pp. 6375–6395. DOI: 10.1029/JB091iB06p06375.

[545] Michael Gurnis and Geoffrey F Davies. "The effect of depth-dependent viscosity on convective mixing in the mantle and the possible survival of primitive mantle". In: *Geophys. Res. Lett.* 13.6 (1986), pp. 541–544. DOI: 10.1029/GL013i006p00541.

[546] Michael Gurnis, Christophe Eloy, and Shijie Zhong. "Free-surface formulation of mantle convection - II. Implication for subduction-zone observables". In: *Geophy. J. Int.* 127.3 (1996), pp. 719–727. DOI: 10.1111/j.1365-246X.1996.tb04050.x.

[547] M-A Gutscher et al. "Thermal modeling of the SW Ryukyu forearc (Taiwan): Implications for the seismogenic zone and the age of the subducting Philippine Sea Plate (Huatung Basin)". In: *Tectonophysics* 692 (2016), pp. 131–142.

[548] Pierre Guyot and John E Dorn. "A critical review of the Peierls mechanism". In: *Canadian Journal of Physics* 45.2 (1967), pp. 983–1016. DOI: 10.1139/p67-073.

[549] B.H. Hager and R.J. O'Connell. "A simple global model of plate dynamics and mantle convection". In: *J. Geophys. Res.: Solid Earth* 86.B6 (1981), pp. 4843–4867. DOI: 10.1029/JB086iB06p04843.

[550] B.H. Hager and R.J. O'Connell. "Subduction zone dip angles and flow driven by plate motion". In: *Tectonophysics* 50 (1978), pp. 111–133. DOI: 10.1016/0040-1951(78)90130-0.

[551] Bradford H Hager. "Subducted slabs and the geoid: Constraints on mantle rheology and flow". In: *J. Geophys. Res.: Solid Earth* 89.B7 (1984), pp. 6003–6015. DOI: 10.1029/JB089iB07p06003.

[552] Bradford H Hager, Robert W Clayton, Mark A Richards, Robert P Comer, and Adam M Dziewonski. "Lower mantle heterogeneity, dynamic topography and the geoid". In: *Nature* 313.6003 (1985), p. 541. DOI: 10.1038/313541a0.

[553] Bradford H Hager and Richard J O'Connell. "Kinematic models of large-scale flow in the Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 84.B3 (1979), pp. 1031–1048. DOI: 10.1029/JB084iB03p01031.

[554] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I*. Springer, 1993.

[555] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, 1991.

[556] A.J. Hale, L. Bourgouin, and H.B. Muehlhaus. "Using the level set method to model endogenous lava dome growth ". In: *J. Geophys. Res.: Solid Earth* 112 (2007), B03213. DOI: 10.1029/2006JB004445.

[557] A.J. Hale, K.-D. Gottschaldt, G. Rosenbaum, L. Bourgouin, M. Bauchy, and Hans Mühlhaus. "Dynamics of slab tear faults: Insights from numerical modelling". In: *Tectonophysics* 483 (2010), pp. 58–70. DOI: 10.1016/j.tecto.2009.05.019.

[558] Andrea Hampel, Jens Lüke, Thomas Krause, and Ralf Hetzel. "Finite-element modelling of glacial isostatic adjustment (GIA): Use of elastic foundations at material boundaries versus the geometrically non-linear formulation". In: *Computers and Geosciences* 122 (2019), pp. 1–14. DOI: 10.1016/j.cageo.2018.08.002.

[559] H. Han. "A finite element approximation of Navier-Stokes equations using nonconforming elements". In: *Journal of Computational Mathematics* 2.1 (1984), pp. 77–88. DOI: xxxx.

[560] J. Handin. "On the Coulomb-Mohr failure criterion". In: *J. Geophys. Res.: Solid Earth* 74.22 (1969), p. 5343.

[561] Emmanuel Hanert, Vincent Legat, and Éric Deleersnijder. "A comparison of three finite elements to solve the linear shallow water equations". In: *Ocean Modelling* 5.1 (2003), pp. 17–35. DOI: 10.1016/S1463-5003(02)00012-4.

[562] P. Hansbo. "A nonconforming rotated Q_1 approximation on tetrahedra". In: *Computer Methods in Applied Mechanics and Engineering* 200 (2011), pp. 1311–1316. DOI: 10.1016/j.cma.2010.11.002.

[563] Peter Hansbo, Mats Larson, and Mats G Larson. "A simple nonconforming bilinear element for the elasticity problem". In: *Trends in Computational Structural Mechanics*. 2001.

[564] D.L. Hansen. "A meshless formulation for geodynamic modeling". In: *J. Geophys. Res.: Solid Earth* 108 (2003). DOI: 10.1029/2003JB002460.

[565] D.L. Hansen and S.B. Nielsen. "Why rifts invert in compression". In: *Tectonophysics* 373 (2003), pp. 5–24.

[566] D.L. Hansen, S.B. Nielsen, and H. Lykke-Andersen. "The post-Triassic evolution of the Sorgenfrei-Tornquist Zone - results from thermo-mechanical modelling". In: *Tectonophysics* 328 (2000), pp. 245–267.

[567] U Hansen and A Ebel. "Experiments with a numerical model related to mantle convection: boundary layer behaviour of small-and large scale flows". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 374–390. DOI: 10.1016/0031-9201(84)90058-X.

[568] U. Hansen and D.A. Yuen. "Effects of depth-dependent thermal expansivity on the interaction of thermal-chemical plumes with a compositional boundary". In: *Phys. Earth. Planet. Inter.* 86.1-3 (1994), pp. 205–221. DOI: 10.1016/0031-9201(94)05069-4.

[569] U. Hansen and D.A. Yuen. "Evolutionary structures in double-diffusive convection in magma chambers". In: *Geophys. Res. Lett.* 14.11 (1987), pp. 1099–1102. DOI: 10.1029/GL014i011p01099.

[570] U. Hansen and D.A. Yuen. "High Rayleigh number regime of temperature-dependent viscosity convection and the Earth's early thermal history". In: *Geophys. Res. Lett.* 20.20 (1993), pp. 2191–2194. DOI: 10.1029/93GL02416.

[571] U. Hansen and D.A. Yuen. "Nonlinear physics of double-diffusive convection in geological systems". In: *Earth Science Reviews* 29.1-4 (1990), pp. 385–399. DOI: 10.1016/0012-8252(90)90050-6.

[572] U. Hansen, D.A. Yuen, and S.E. Kroening. "Mass and Heat Transport in Strongly Time-Dependent Thermal Convection at Infinite Prandtl Number". In: *Geophysical & Astrophysical Fluid Dynamics* 63.1-4 (1992), pp. 67–89. DOI: 10.1080/03091929208228278.

[573] U. Hansen, D.A. Yuen, and A.V. Malevsky. "Comparison of steady-state and strongly chaotic thermal convection at high Rayleigh number". In: *Physical Review A* 46.8 (1992), pp. 4742–4754. DOI: 10.1103/PhysRevA.46.4742.

[574] Ulrich Hansen and Adolf Ebel. "Time-dependent thermal convection-a possible explanation for a multiscale flow in the Earth's mantle". In: *Geophy. J. Int.* 94.2 (1988), pp. 181–191. DOI: 10.1111/j.1365-246X.1988.tb05895.x.

[575] Ulrich Hansen and David A Yuen. "Dynamical influences from thermal-chemical instabilities at the core-mantle boundary". In: *Geophys. Res. Lett.* 16.7 (1989), pp. 629–632. DOI: 10.1029/GL016i007p00629.

[576] Ulrich Hansen, David A Yuen, SE Kroening, and TB Larsen. "Dynamical consequences of depth-dependent thermal expansivity and viscosity on mantle circulations and thermal structure". In: *Phys. Earth. Planet. Inter.* 77.3-4 (1993), pp. 205–223. DOI: 10.1016/0031-9201(93)90099-U.

[577] Ulrich Hansen, David A Yuen, and Sherri E Kroening. "Effects of depth-dependent thermal expansivity on mantle circulations and lateral thermal anomalies". In: *Geophys. Res. Lett.* 18.7 (1991), pp. 1261–1264. DOI: 10.1029/91GL01288.

[578] L. Hanyk, J. Moser, D.A. Yuen, and C. Matyska. "Time-domain approach for the transient responses in stratified viscoelastic Earth models". In: *Geophys. Res. Lett.* 22.10 (1995), pp. 1285–1288. DOI: 10.1029/95GL01087.

[579] Helmut Harder. "Numerical simulation of thermal convection with Maxwellian viscoelasticity". In: *Journal of non-newtonian fluid mechanics* 39.1 (1991), pp. 67–88. DOI: 10.1016/0377-0257(91)80004-4.

[580] Helmut Harder. "Phase transitions and the three-dimensional planform of thermal convection in the Martian mantle". In: *J. Geophys. Res.: Planets* 103.E7 (1998), pp. 16775–16797. DOI: 10.1029/98JE01543.

[581] Doug P Hardin, Edward B Saff, et al. "Discretizing manifolds via minimum energy points". In: *Notices of the AMS* 51.10 (2004), pp. 1186–1194. DOI: xxxx.

[582] F.H. Harlow and J.E. Welch. "Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface". In: *The physics of fluids* 8.12 (1965), p. 2182.

[583] Nathan J Harris, Steven E Benzley, and Steven J Owen. "Conformal Refinement of All-Hexahedral Element Meshes Based on Multiple Twist Plane Insertion". In: *IMR*. 2004, pp. 157–168.

[584] E.H. Hartz and Y.Y. Podlachikov. "Toasting the jelly sandwich: The effect of shear heating on lithospheric geotherms and strength ". In: *Geology* 36.4 (2008), pp. 331–334.

[585] N.A. Haskell. "The Motion of a Viscous Fluid Under a Surface Load". In: *Physics* 6 (1935), pp. 265–269. DOI: 10.1063/1.1745329.

[586] R. Hassan, N. Flament, M. Gurnis, D.J. Bower, and D. Múller. "Provenance of plumes in global convection models". In: *Geochem. Geophys. Geosyst.* 16 (2015), pp. 1465–1489. DOI: 10.1002/2015GC005751.

[587] R. Hassani and J. Chéry. "Anelasticity explains topography associated with Basin and Range normal faulting". In: *Geology* 24.12 (1996), pp. 1095–1098. DOI: 10.1130/0091-7613(1996)024<1095:AETAWB>2.3.CO;2.

[588] Hiroshi Hayashi and Akira Kageyama. "Yin–Yang–Zhong grid: An overset grid system for a sphere". In: *J. Comp. Phys.* 305 (2016), pp. 895–905. DOI: 10.1016/j.jcp.2015.11.016.

[589] Y. He, E.G. Puckett, and M.I. Billen. "A discontinuous Galerkin method with a bound preserving limiter for the advection of non-diffusive fields in solid Earth geodynamics". In: *Phys. Earth. Planet. Inter.* 263 (2017), pp. 23–37.

[590] B. Heck and K. Seitz. "A comparison of the tesseroid, prism and point-mass approaches for mass reductions in gravity field modelling". In: *J. Geodesy* 81 (2007), pp. 121–136. DOI: 10.1007/s00190-006-0094-0.

[591] Otto M Heeres, Akke SJ Suiker, and René de Borst. "A comparison between the Perzyna viscoplastic model and the consistency viscoplastic model". In: *European Journal of Mechanics-A/Solids* 21.1 (2002), pp. 1–12. DOI: 10.1016/S0997-7538(01)01188-3.

[592] Juan C Heinrich, Peter S Huyakorn, Olgierd C Zienkiewicz, and AR0353 Mitchell. "An 'upwind' finite element scheme for two-dimensional convective transport equation". In: *Int. J. Num. Meth. Eng.* 11.1 (1977), pp. 131–143. DOI: 10.1002/nme.1620110113.

[593] T. Heister, J. Dannberg, R. Gassmöller, and W. Bangerth. "High Accuracy Mantle Convection Simulation through Modern Numerical Methods. II: Realistic Models and Problems". In: *Geophy. J. Int.* 210.2 (2017), pp. 833–851. DOI: 10.1093/gji/ggx195.

[594] Christian Helanow and Josefin Ahlkrona. "Stabilized equal low-order finite elements in ice sheet modeling–accuracy and robustness". In: *Computational Geosciences* 22.4 (2018), pp. 951–974. DOI: 10.1007/s10596-017-9713-5.

[595] Heinrich Hencky. "Zur Theorie plastischer Deformationen und der hierdurch im Material hervorgerufenen Nachspannungen". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 4.4 (1924), pp. 323–334. DOI: 10.1002/zamm.19240040405.

[596] P. Hénon, P. Ramet, and J. Roman. "PaStiX: A High-Performance Parallel Direct Solver for Sparse Symmetric Definite Systems". In: *Parallel Computing* 28.2 (2002), pp. 301–321.

[597] K.J. Bathe, ed. *Three-dimensional spherical shell convection at infinite Prandtl number using the 'cubed sphere' method.* Proceedings Second MIT Conference on Compurational Fluid and Solid Mechanics June 17-20, 2003. 2003, pp. 931–933. DOI: xxxx.

[598] John W Hernlund, Christine Thomas, and Paul J Tackley. "A doubling of the post-perovskite phase boundary and structure of the Earth's lowermost mantle". In: *Nature* 434.7035 (2005), p. 882. DOI: 10.1038/nature03472.

[599] Marco Herwegh, T Poulet, Ali Karrech, and Klaus Regenauer-Lieb. "From transient to steady state deformation and grain size: A thermodynamic approach using elasto-visco-plastic numerical modeling". In: *J. Geophys. Res.: Solid Earth* 119.2 (2014), pp. 900–918. DOI: 10.1002/2013JB010701.

[600] M.R. Hestenes and E. Stiefel. "Methods of Conjugate Gradients for Solving Linear Systems". In: *Journal of Research of the National Bureau of Standards* 49.6 (1952), pp. 409–436.

[601] J.S. Hesthaven and T. Warbuton. "Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications". In: *Texts in Applied Mathematics*. Ed. by J.E. Marsden, L. Sirovich, and S.S. Antman. Springer, 2008.

[602] J.M. Hewitt, D.P. McKenzie, and N.O. Weiss. "Dissipative heating in convective flows". In: *J. Fluid Mech.* 68.4 (1975), pp. 721–738. DOI: 10.1017/S002211207500119X.

[603] N. Hilairet et al. "High-Pressure creep of Serpentine, interseismic deformation, and initiation of subduction". In: *Science* 318 (2007), pp. 1910–1913.

[604] B. Hillebrand, C. Thieulot, T. Geenen, A. P. van den Berg, and W. Spakman. "Using the level set method in geodynamical modeling of multi-material flows and Earth's free surface". In: *Solid Earth* 5.2 (2014), pp. 1087–1098. DOI: `10.5194/se-5-1087-2014`.

[605] RD van der Hilst, MV e Hoop, P Wang, S-H Shim, P Ma, and L Tenorio. "Seismostratigraphy and thermal structure of Earth's core-mantle boundary region". In: *science* 315.5820 (2007), pp. 1813–1817. DOI: `10.1126/science.1137867`.

[606] Kei Hirose, John Brodholt, Thorne Lay, David A Yuen, et al. "An Introduction to Post-Perovskite: The Last Mantle Phase Transition". In: *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION* 174 (2007), p. 1.

[607] C.W Hirt and B.D Nichols. "Volume of fluid (VOF) method for the dynamics of free boundaries". In: *J. Computational Physics* 39.1 (1981), pp. 201–225. DOI: `10.1016/0021-9991(81)90145-5`.

[608] C.W. Hirt, A. Amsden, and J.L. Cook. "An Arbitrary Lagrangian-Eulerian computing method for all flow speeds". In: *J. Comp. Phys.* 14 (1974), pp. 227–253. DOI: `10.1016/0021-9991(74)90051-5`.

[609] Christian Hirt, SJ Claessens, Michael Kuhn, and WE Featherstone. "Kilometer-resolution gravity field of Mars: MGM2011". In: *Planetary and Space Science* 67.1 (2012), pp. 147–154. DOI: `10.1016/j.pss.2012.02.006`.

[610] G. Hirth and D. L. Kohlstedt. "Water in the oceanic upper mantle: Implications for rheology, melt extraction and the evolution of the lithosphere". In: *Earth Planet. Sci. Lett.* 144 (1996), pp. 93–108. DOI: `10.1016/0012-821X(96)00154-9`.

[611] G. Hirth and D.L. Kohlstedt. "Rheology of the upper mantle and the mantle wedge: A view from the experimentalists". In: *in Inside the Subduction Factory, ed. J. Eiler, Geophysical Monograph American Geophysical Union, Washington, D.C.* 138 (2003), pp. 83–105.

[612] BE Hobbs, H-B Mühlhaus, and A Ord. "Instability, softening and localization of deformation". In: *Geological Society, London, Special Publications* 54.1 (1990), pp. 143–165. DOI: `10.1144/GSL.SP.1990.054.01.15`.

[613] BE Hobbs and A Ord. "Numerical simulation of shear band formation in a frictional-dilational material". In: *Ingenieur-Archiv* 59.3 (1989), pp. 209–220. DOI: `10.1007/BF00532251`.

[614] NRA Hoffman and DP McKenzie. "The destruction of geochemical heterogeneities by differential fluid motions during mantle convection". In: *Geophy. J. Int.* 82.2 (1985), pp. 163–206. DOI: `10.1111/j.1365-246X.1985.tb05134.x`.

[615] Klaus A Hoffmann and Steve T Chiang. "Computational fluid dynamics for engineers, Vol. 1". In: *Wichita, Engineering Education System* (1993), pp. 124–137.

[616] A.M. Hofmeister and D.A. Yuen. "Critical phenomena in thermal conductivity: Implications for lower mantle dynamics". In: *Journal of Geodynamics* 44.3-5 (2007), pp. 186–199. DOI: `10.1016/j.jog.2007.03.002`.

[617] AJ Hogg and GP Matson. "Slumps of viscoplastic fluids on slopes". In: *Journal of Non-Newtonian Fluid Mechanics* 158.1-3 (2009), pp. 101–112. DOI: `10.1016/j.jnnfm.2008.07.003`.

[618] John C Holden and Peter Vogt. "Graphic solutions to problems of plumacy". In: *Eos, Transactions American Geophysical Union* 58.7 (1977), pp. 573–580. DOI: `10.1029/EO058i007p00573`.

[619] Bjørn Holmedal. "Spin and vorticity with vanishing rigid-body rotation during shear in continuum mechanics". In: *Journal of the Mechanics and Physics of Solids* 137 (2020), p. 103835. DOI: `10.1016/j.jmps.2019.103835`.

[620] Satoru Honda. "Thermal structure beneath Tohoku, northeast Japan". In: *Tectonophysics* 112.1-4 (1985), pp. 69–102. DOI: 10.1016/0040-1951(85)90173-8.

[621] Seungwoo Hong, Kyungsoo Kim, and Sungyun Lee. "Modified cross-grid finite elements for the stokes problem". In: *Applied mathematics letters* 16.1 (2003), pp. 59–64. DOI: 10.1016/S0893-9659(02)00144-1.

[622] P. Hood and C. Taylor. "Navier-Stokes equations using mixed interpolation". In: *Finite element methods in flow problems*. Ed. by J.T. Oden, R.H. Gallagher, O.C. Zienkiewicz, and C. Taylor. University of Alabama, 1974.

[623] André Horbach, Hans-Peter Bunge, and Jens Oeser. "The adjoint method in geodynamics: derivation from a general operator formulation and application to the initial condition problem in a high resolution mantle circulation model". In: *GEM-International Journal on Geomathematics* 5.2 (2014), pp. 163–194.

[624] André Horbach, Marcus Mohr, and Hans-Peter Bunge. "A semi-analytic accuracy benchmark for Stokes flow in 3-D spherical mantle convection codes". In: *GEM-International Journal on Geomathematics* 11.1 (2020), pp. 1–35. DOI: 10.1007/s13137-019-0137-3.

[625] C.O. Horgan. "Korn's inequalities and their applications in continuum mechanics". In: *SIAM Review* 37.4 (1995), pp. 491–511. DOI: 10.1137/1037123.

[626] GA Houseman. "Boundary conditions and efficient solution algorithms for the potential function formulation of the 3-D viscous flow equations". In: *Geophy. J. Int.* 100.1 (1990), pp. 33–38. DOI: 10.1111/j.1365-246X.1990.tb04565.x.

[627] GA Houseman. "The thermal structure of mantle plumes: axisymmetric or triple-junction?" In: *Geophy. J. Int.* 102.1 (1990), pp. 15–24. DOI: 10.1111/j.1365-246X.1990.tb00527.x.

[628] Greg A Houseman, D Po McKenzie, and Peter Molnar. "Convective instability of a thickened boundary layer and its relevance for the thermal evolution of continental convergent belts". In: *J. Geophys. Res.: Solid Earth* 86.B7 (1981), pp. 6115–6132. DOI: 10.1029/JB086iB07p06115.

[629] Gregory Houseman and Philip England. "Finite strain calculations of continental deformation: 1. Method and general results for convergent zones". In: *J. Geophys. Res.: Solid Earth* 91.B3 (1986), pp. 3651–3663. DOI: 10.1029/JB091iB03p03651.

[630] Albert T Hsui. "Numerical simulation of finite-amplitude thermal convection with large viscosity variation in axisymmetric spherical geometry: effect of mechanical boundary conditions". In: *Tectonophysics* 50.2-3 (1978), pp. 147–162. DOI: 10.1016/0040-1951(78)90132-4.

[631] Albert T Hsui, Tang Xiao-Ming, and M Nafi Toksöz. "On the dip angle of subducting plates". In: *Tectonophysics* 179.3-4 (1990), pp. 163–175.

[632] Chongyu Hua. "An inverse transformation for quadrilateral isoparametric elements: analysis and application". In: *Finite elements in analysis and design* 7.2 (1990), pp. 159–166. DOI: 10.1016/0168-874X(90)90007-2.

[633] Yunqing Huang and Shangyou Zhang. "A lowest order divergence-free finite element on rectangular grids". In: *Frontiers of Mathematics in China* 6 (2011), pp. 253–270. DOI: 10.1007/s11464-011-0094-0.

[634] Ch. Huber, A. Parmigiani, B. Chopard, M. Manga, and O. Bachmann. "Lattice Boltzmann model for melting with natural convection". In: *International Journal of Heat and Fluid Flow* 29 (200), pp. 1469–1480. DOI: 10.1016/j.ijheatfluidflow.2008.05.002.

[635] Aurélia Hubert-Ferrari, Geoffrey King, Isabelle Manighetti, Rolando Armijo, Bertrand Meyer, and Paul Tapponnier. "Long-term elasticity in the continental lithosphere; modelling the Aden Ridge propagation and the Anatolian extrusion process". In: *Geophy. J. Int.* 153.1 (2003), pp. 111–132. DOI: 10.1046/j.1365-246X.2003.01872.x.

[636] A. Huerta and W.K. Liu. "Viscous flow with large free surface motion". In: *Computer Methods in Applied Mechanics and Engineering* 69 (1988), pp. 277–324.

[637] T.J.R. Hughes. *The Finite Element Method. Linear Static and Dynamic Finite Element Analysis*. Dover Publications, Inc., 2000. ISBN: 0-486-41181-8.

[638] T.J.R. Hughes and A. Brooks. "A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: application to the streamline-upwind procedure". In: *Finite Elements in Fluids* 4 (1982), pp. 47–65.

[639] T.J.R. Hughes, L.P. Franca, and M. Balestra. "A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations". In: *Computer Methods in Applied Mechanics and Engineering* 59.1 (1986), pp. 85–99.

[640] T.J.R. Hughes, L.P. Franca, and M. Balestra. "A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces". In: *Computer Methods in Applied Mechanics and Engineering* 65 (1987), pp. 85–96.

[641] T.J.R. Hughes, W.K. Liu, and A. Brooks. "Finite element analysis of Incompressible viscous flows by the penalty function formulation". In: *J. Comp. Phys.* 30 (1979), pp. 1–60. DOI: 10.1016/0021-9991(79)90086-X.

[642] Thomas JR Hughes, Wing Kam Liu, and Thomas K Zimmermann. "Lagrangian-Eulerian finite element formulation for incompressible viscous flows". In: *Computer Methods in Applied Mechanics and Engineering* 29.3 (1981), pp. 329–349. DOI: 10.1016/0045-7825(81)90049-9.

[643] Thomas JR Hughes, Michel Mallet, and Akira Mizukami. "A new finite element formulation for computational fluid dynamics: II. Beyond SUPG". In: *Computer Methods in Applied Mechanics and Engineering* 54.3 (1986), pp. 341–355. DOI: 10.1016/0045-7825(86)90110-6.

[644] Hoon Huh, Choong Ho Lee, and Wei H. Yang. "A general algorithm for plastic flow simulation by finite element limit analysis". In: *International Journal of Solids and Structures* 36 (1999), pp. 1193–1207. DOI: 10.1016/S0020-7683(97)00347-8.

[645] R. Huismans and C. Beaumont. "Depth-dependent extension, two-stage breakup and cratonic underplating at rifted margins". In: *Nature* 473 (2011), pp. 74–79. DOI: 10.1038/nature09988.

[646] R. S. Huismans and C. Beaumont. "Symmetric and asymmetric lithospheric extension: Relative effects of frictional-plastic and viscous strain softening". In: *J. Geophys. Res.: Solid Earth* 108 (B10).2496 (2003). DOI: 10.1029/2002JB002026.

[647] R.S. Huismans and C. Beaumont. "Roles of lithospheric strain softening and heterogeneity in determining the geometry of rifts and continental margins". In: *Imaging, Mapping and Modelling Continental Lithosphere Extension and Breakup*. Vol. 282. Geological Society, London, Special Publications, 2007, pp. 111–138. DOI: 10.1144/SP282.6.

[648] R.S. Huismans, S.J.H. Buiter, and C. Beaumont. "Effect of plastic-viscous layering and strain softening on mode selection during lithospheric extension". In: *J. Geophys. Res.: Solid Earth* 110 (2005), B02406. DOI: 10.1029/2004JB003114.

[649] C. Hüttig and K. Stemmer. "The spiral grid: A new approach to discretize the sphere and its application to mantle convection". In: *Geochem. Geophys. Geosyst.* 9.2 (2008). DOI: 10.1029/2007GC001581.

[650] G. Iaffaldano and H.-P. Bunge. "Relating rapid plate-motion variations to plate-boundary forces in global coupled models of the mantle/lithosphere system: Effects of topography and friction". In: *Tectonophysics* 474.1-2 (2009), pp. 393–404. DOI: 10.1016/j.tecto.2008.10.035.

[651] S. Idelsohn, M. Storti, and N. Nigro. "Stability analysis of mixed finite element formulations with special mention of equal-order interpolations". In: *Int. J. Num. Meth. Fluids* 20 (1995), pp. 1003–1022. DOI: 10.1002/fld.1650200819.

[652] Ponsuganth Ilangovan, Eugenio D'Ascoli, Nils Kohl, and Marcus Mohr. "Numerical Studies on Coupled Stokes-Transport Systems for Mantle Convection". In: *International Conference on Computational Science.* 2024, pp. 288–302. DOI: 10.1007/978-3-031-63759-9_33.

[653] F. Ilinca and D. Pelletier. "Computation of accurate nodal derivatives of finite element solutions: The finite node displacement method". In: *Int. J. Num. Meth. Eng.* 71 (2007), pp. 1181–1207. DOI: 10.1002/nme.1979.

[654] Wadi H Imseeh and Khalid A Alshibli. "3D finite element modelling of force transmission and particle fracture of sand". In: *Computers and Geotechnics* 94 (2018), pp. 184–195. DOI: 10.1016/j.compgeo.2017.09.008.

[655] Computational Inelasticity. 1998.

[656] Tobin Isaac, Georg Stadler, and Omar Ghattas. "Solution of nonlinear Stokes equations discretized by high-order finite elements on nonconforming and anisotropic meshes, with application to ice sheet dynamics". In: *SIAM Journal on Scientific Computing* 37.6 (2015), B804–B833. DOI: 10.1137/140974407.

[657] Hideyuki Ishii, Yoshihisa Sakurai, and Takeo Maruyama. "Effect of soccer shoe upper on ball behaviour in curve kicks". In: *Scientific Reports* 4 (2014), p. 6067. DOI: 10.1038/srep06067.

[658] A Ismail-Zadeh, A Korotkii, G Schubert, and I Tsepelev. "Quasi-reversibility method for data assimilation in models of mantle dynamics". In: *Geophy. J. Int.* 170.3 (2007), pp. 1381–1398. DOI: 10.1111/j.1365-246X.2007.03496.x.

[659] Alik Ismail-Zadeh and Paul Tackley. *Computational Methods for Geodynamics.* Cambridge University Press, 2010.

[660] J. Ita and S.D. King. "Sensitivity of convection with an endothermic phase change to the form of governing equations, initial conditions, boundary conditions, and equation of state". In: *J. Geophys. Res.: Solid Earth* 99.B8 (1994), pp. 15, 919–15, 938. DOI: 10.1029/94JB00852.

[661] Y. Iwase. "Three-dimensional infinite Prandtl number convection in a spherical shell with temperature-dependent viscosity". In: *J. Geomag. Geoelectr.* 48 (1996), pp. 1499–1514. DOI: 10.5636/jgg.48.1499.

[662] Michel HG Jacobs and Bernard HWS de Jong. "Placing constraints on phase equilibria and thermophysical properties in the system MgO–SiO2 by a thermodynamically consistent vibrational method". In: *Geochimica et Cosmochimica Acta* 71.14 (2007), pp. 3630–3655. DOI: 10.1016/j.gca.2007.05.010.

[663] Wolfgang R Jacoby and Harro Schmeling. "Convection experiments and the driving mechanism". In: *Geologische Rundschau* 70.1 (1981), pp. 207–230. DOI: 10.1007/BF01764323.

[664] Wolfgang R Jacoby and Harro Schmeling. "On the effects of the lithosphere on mantle convection and evolution". In: *Phys. Earth. Planet. Inter.* 29.3-4 (1982), pp. 305–319. DOI: 10.1016/0031-9201(82)90019-X.

[665] J.C. Jaeger. *Elasticity, Fracture and Flow*. John Wiley and Sons, Inc., 1969.

[666] S. Jammes and R.S. Huismans. "Structural styles of mountain building: Controls of lithospheric rheologic stratification and extensional inheritance". In: *J. Geophys. Res.: Solid Earth* 117.B10 (2012). DOI: 10.1029/2012JB009376.

[667] Gang-Won Jang, Sangkeun Lee, Yoon Young Kim, and Dongwoo Sheen. "Topology optimization using non-conforming finite elements: three-dimensional case". In: *Int. J. Num. Meth. Eng.* 63.6 (2005), pp. 859–875. DOI: 10.1002/nme.1302.

[668] O. Jaoul, J. Tullis, and A. Kronenberg. "The effect of varying water contents on the creep behavior of Heavitree quartzite". In: *J. Geophys. Res.: Solid Earth* 89.B6 (1984). DOI: 10.1029/JB089iB06p04298.

[669] G.T. Jarvis. "Effects of curvature on two-dimensional models of mantle convection: cylindrical polar coordinates". In: *J. Geophys. Res.: Solid Earth* 98.B3 (1993), pp. 4477–4485. DOI: 10.1029/92JB02117.

[670] Gary T Jarvis and WR Peltier. "Effects of lithospheric rigidity on ocean floor bathymetry and heat flow". In: *Geophys. Res. Lett.* 8.8 (1981), pp. 857–860. DOI: 10.1029/GL008i008p00857.

[671] Gary T Jarvis and WR Peltier. "Mantle convection as a boundary layer phenomenon". In: *Geophy. J. Int.* 68.2 (1982), pp. 389–427. DOI: 10.1111/j.1365-246X.1982.tb04907.x.

[672] Gary T. Jarvis and Dan P. McKenzie. "Convection in a compressible fluid with infinite Prandtl number". In: *Journal of Fluid Mechanics* 96.3 (1980), pp. 515–583. DOI: 10.1017/S002211208000225X.

[673] GT Jarvis. "Time-dependent convection in the Earth's mantle". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 305–327. DOI: 10.1016/0031-9201(84)90054-2.

[674] GT Jarvis. "Two-dimensional numerical models of mantle convection". In: *Advances in geophysics*. Vol. 33. 1991, pp. 1–80. DOI: 10.1016/S0065-2687(08)60440-9.

[675] GT Jarvis and WR Peltier. "Lateral heterogeneity in the convecting mantle". In: *J. Geophys. Res.: Solid Earth* 91.B1 (1986), pp. 435–451. DOI: 10.1029/JB091iB01p00435.

[676] GT Jarvis and WR Peltier. "Low-wavenumber signatures of time-dependent mantle convection". In: *Phys. Earth. Planet. Inter.* 59.3 (1990), pp. 182–194. DOI: 10.1016/0031-9201(90)90224-L.

[677] C. Jaupart and J.-C. Mareschal. *Heat Generation and Transport in the Earth*. Cambridge, 2011.

[678] P. Jay, A. Magnin, and J.-M. Piau. "Viscoplastic Fluid Flow Through a Sudden Axisymmetric Expansion". In: *AIChE Journal* 47.10 (2001), pp. 2155–2166. DOI: 10.1002/aic.690471004.

[679] Raymond Jeanloz and S Morris. "Is the mantle geotherm subadiabatic?" In: *Geophys. Res. Lett.* 14.4 (1987), pp. 335–338. DOI: 10.1029/GL014i004p00335.

[680] Raymond Jeanloz and S Morris. "Temperature distribution in the crust and mantle". In: *Annual Review of Earth and Planetary Sciences* 14.1 (1986), pp. 377–415. DOI: xxxx.

[681] Eleanor W Jenkins, Volker John, Alexander Linke, and Leo G Rebholz. "On the parameter choice in grad-div stabilization for the Stokes equations". In: *Advances in Computational Mathematics* 40.2 (2014), pp. 491–516. DOI: 10.1007/s10444-013-9316-1.

[682] P. Jenny, S.B. Pope, M. Muradoglu, and D.A. Caughey. "A Hybrid Algorithm for the Joint PDF Equation of Turbulent Reactive Flows". In: *J. Comp. Phys.* 166 (2001), pp. 218–252. DOI: 10.1006/jcph.2000.6646.

[683] Youngmok Jeon, Hyun Nam, Dongwoo Sheen, and Kwangshin Shim. "A class of nonparametric DSSY nonconforming quadrilateral elements". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 47.6 (2013), pp. 1783–1796. DOI: 10.1051/m2an/2013088.

[684] L. Jiang, J. Liu, J. Zhang, and Z. Feng. "Analytic Expressions for the Gravity Gradient Tensor of 3D Prisms with Depth-Dependent Density". In: *Surv. Geophys.* 39 (2018), pp. 337–363. DOI: 10.1007/s10712-017-9455-x.

[685] Liqing Jiao, Paul Tapponnier, Frédéric-victor Donzé, Luc Scholtès, Yves Gaudemer, and Xiwei Xu. "Discrete element modeling of southeast Asia's 3D lithospheric deformation during the Indian collision". In: *J. Geophys. Res.: Solid Earth* 128.1 (2023). DOI: 10.1029/2022JB025578.

[686] Stephen Jiménez, Ravindra Duddu, and Jeremy Bassis. "An updated-Lagrangian damage mechanics formulation for modeling the creeping flow and fracture of ice sheets". In: *Computer Methods in Applied Mechanics and Engineering* 313 (2017), pp. 406–432. DOI: 10.1016/j.cma.2016.09.034.

[687] Volker John. "A posteriori L^2-error estimates for the nonconforming P_1/P_0-finite element discretization of the Stokes equations". In: *Journal of computational and applied mathematics* 96.2 (1998), pp. 99–116. DOI: 10.1016/S0377-0427(98)00095-8.

[688] Volker John. *Finite Element Methods for Incompressible Flow Problems.* Springer, 2016. ISBN: 978-3-319-45749-9. DOI: 10.1007/978-3-319-45750-5.

[689] Volker John. "Higher order finite element methods and multigrid solvers in a benchmark problem for the 3D Navier–Stokes equations". In: *Int. J. Num. Meth. Fluids* 40.6 (2002), pp. 775–798. DOI: 10.1002/d.377.

[690] Volker John, Kristine Kaiser, and Julia Novo. "Finite element methods for the incompressible Stokes equations with variable viscosity". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 96.2 (2016), pp. 205–216. DOI: 10.1002/zamm.201400291.

[691] Volker John, Songul Kaya, and Julia Novo. *Finite Element Error Analysis Of A Mantle Convection Model.* Vol. 15. 4–5. 2018, pp. 677–698. DOI: xxx.

[692] Volker John and Petr Knobloch. "On discontinuity-capturing methods for convection-diffusion equations". In: *Numerical mathematics and advanced applications.* 2006, pp. 336–344. DOI: 10.1007/978-3-540-34288-5_27.

[693] Volker John, Alexander Linke, Christian Merdon, Michael Neilan, and Leo G Rebholz. "On the divergence constraint in mixed finite element methods for incompressible flows". In: *SIAM review* 59.3 (2017), pp. 492–544. DOI: 10.1137/15M1047696.

[694] L. Jolivet, P. Davy, and P. Cobbold. "Right-lateral shear along the Northwest Pacific margin and the India-Eurasia collision". In: *Tectonics* 9.6 (1990), pp. 1409–1419. DOI: 10.1029/TC009i006p01409.

[695] MS Joun and MC Lee. "Quadrilateral finite-element generation and mesh quality control for metal forming simulation". In: *Int. J. Num. Meth. Eng.* 40.21 (1997), pp. 4059–4075. DOI: 10.1002/(SICI)1097-0207(19971115)40:21<4059::AID-NME249>3.0.CO;2-E.

[696] Guillaume Jouvet and Jacques Rappaz. "Analysis and finite element approximation of a nonlinear stationary Stokes problem arising in glaciology". In: *Advances in Numerical Analysis* 2011 (2011). DOI: 10.1155/2011/164581.

[697] M Jull and D McKenzie. "The effect of deglaciation on mantle melting beneath Iceland". In: *J. Geophys. Res.: Solid Earth* 101.B10 (1996), pp. 21815–21828. DOI: 10.1029/96JB01308.

[698] Mikhail K Kaban, Alexey G Petrunin, Harro Schmeling, and Meysam Shahraki. "Effect of decoupling of lithospheric plates on the observed geoid". In: *Surveys in Geophysics* 35.6 (2014), pp. 1361–1373. DOI: 10.1007/s10712-014-9281-3.

[699] L.M. Kachanov. *Fundamentals of the Theory of Plasticity*. Dover Publications, Inc., 2004. ISBN: 0-486-43583-0.

[700] B.J. Kadlec, G.A. Dorn, H.M. Tufo, and D.A. Yuen. "Interactive 3-D computation of fault surfaces using level sets". In: *Visual Geosciences* 13.1 (2008), pp. 133–138. DOI: 10.1007/s10069-008-0016-9.

[701] A. Kageyama and T. Sato. ""Yin-Yang grid": An overset grid in spherical geometry". In: *Geochem. Geophys. Geosyst.* 5.9 (2004), 10.1029/2004GC000734.

[702] Juhani Kakkuri. "Fennoscandian Land Uplift: Past, Present and Future". In: *From the Earth's Core to Outer Space*. Ed. by Ilmari Haapala. Berlin, Heidelberg, 2012, pp. 127–136. ISBN: 978-3-642-25550-2. DOI: 10.1007/978-3-642-25550-2_8.

[703] M. Kameyama, A. Kageyamab, and T. Sato. "Multigrid-based simulation code for mantle convection in spherical shell using Yin-Yang grid". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 19–32.

[704] M. Kameyama, D.A. Yuen, and S.-I. Karato. "Thermal-mechanical effects of low-temperature plasticity (the Peierls mechanism) on the deformation of a viscoelastic shear zone". In: *Earth Planet. Sci. Lett.* 168 (1999), pp. 159–172.

[705] Masanori Kameyama. "ACuTEMan: A multigrid-based mantle convection simulation code and its optimization to the Earth Simulator". In: *J. Earth Simulator* 4 (2005), pp. 2–10. DOI: xxxx.

[706] Masanori Kameyama, Akira Kageyama, and Tetsuya Sato. "Multigrid iterative algorithm using pseudo-compressibility for three-dimensional mantle convection with strongly variable viscosity". In: *J. Comp. Phys.* 206.1 (2005), pp. 162–181. DOI: 10.1016/j.jcp.2004.11.030.

[707] Takeo Kaneko, Tomoeki Nakakuki, and Hikaru Iwamori. "Mechanical coupling of the motion of the surface plate and the lower mantle slab: Effects of viscosity hill, yield strength, and depth-dependent thermal expansivity". In: *Phys. Earth. Planet. Inter.* 294 (2019), p. 106274. DOI: 10.1016/j.pepi.2019.106274.

[708] G. Kanschat. *Discontinuous Galerkin Methods for Viscous Incompressible Flow*. Teubner Research, 2007.

[709] G. Kanschat. "Divergence-free discontinuous Galerkin schemes for the Stokes equations and the MAC scheme". In: *Int. J. Num. Meth. Fluids* 56 (2008), pp. 941–950. DOI: 10.1002/fld.1566.

[710] Guido Kanschat. *Mixed Finite Element Methods*. 2017.

[711] Elias Karabelas, Gundolf Haase, Gernot Plank, and Christoph M Augustin. "Versatile stabilized finite element formulations for nearly and fully incompressible solid mechanics". In: *Computational mechanics* 65.1 (2020), pp. 193–215. DOI: 10.1007/s00466-019-01760-w.

[712] S. Karato, M.R. Riedel, and D.A. Yuen. "Rheological structure and deformation of subducted slabs in the mantle transition zone: implications for mantle circulation and deep earthquakes". In: *Phys. Earth. Planet. Inter.* 127 (2001), pp. 83–108. DOI: 10.1016/S0031-9201(01)00223-0.

[713] S.-I. Karato and H. Jung. "Effects of pressure on high-temperature dislocation creep in olivine". In: *Philosophical Magazine* 83.3 (2003), pp. 401–414. DOI: 10.1080/0141861021000025829.

[714] S.-I. Karato and P. Wu. "Rheology of the Upper Mantle: A synthesis". In: *Science* 260 (1993), pp. 771–778. DOI: 10.1126/science.260.5109.771.

[715] Shun-Ichiro Karato. *Deformation of Earth Materials*. Cambridge University Press, 2008. ISBN: 978-0-521-84404-8.

[716] I. Katayama and S. Karato. "Low-temperature, high-stress deformation of olivine under water-saturated conditions". In: *Phys. Earth. Planet. Inter.* 168.3-4 (2008), pp. 125–133.

[717] Tomoo Katsura, Akira Yoneda, Daisuke Yamazaki, Takashi Yoshino, and Eiji Ito. "Adiabatic temperature profile in the mantle". In: *Phys. Earth. Planet. Inter.* 183.1-2 (2010), pp. 212–218. DOI: 10.1016/j.pepi.2010.07.001.

[718] Rafael Katzman, Uri S ten Brink, and Jian Lin. "Three-dimensional modeling of pull-apart basins: Implications for the tectonics of the Dead Sea Basin". In: *J. Geophys. Res.: Solid Earth* 100.B4 (1995), pp. 6295–6312. DOI: 10.1029/94JB03101.

[719] B. Kaus. "Modelling approaches to geodynamic processes". PhD thesis. ETH Zurich, 2005.

[720] B.J.P. Kaus. "Factors that control the angle of shear bands in geodynamic numerical models of brittle deformation". In: *Tectonophysics* 484 (2010), pp. 36–47. DOI: 10.1016/j.tecto.2009.08.042.

[721] B.J.P. Kaus and T.W. Becker. "Effects of elasticity on the Rayleigh-Taylor instability: implications for large-scale geodynamics". In: *Geophy. J. Int.* 168.843–862 (2007). DOI: 10.1111/j.1365-246X.2006.03201.x.

[722] B.J.P. Kaus, H. Mühlhaus, and D.A. May. "A stabilization algorithm for geodynamic numerical simulations with a free surface". In: *Phys. Earth. Planet. Inter.* 181 (2010), pp. 12–20. DOI: 10.1016/j.pepi.2010.04.007.

[723] B.J.P. Kaus and Y.Y. Podlachikov. "Initiation of localized shear zones in viscoelastoplastic rocks". In: *J. Geophys. Res.: Solid Earth* 111.B04412 (2006). DOI: 10.1029/2005JB003652.

[724] B.J.P. Kaus et al. "Forward and Inverse Modelling of Lithospheric Deformation on Geological Timescales". In: *NIC Symposium 2016* (2016), pp. 299–307.

[725] M. Kawaguti. "Numerical solution of the Navier-Stokes equations for the flow in a two-dimensional cavity". In: *Journal of the Physical Society of Japan* 16.12 (1961), pp. 2307–2315.

[726] Takaaki Kawazoe, Shun-ichiro Karato, Kazuhiko Otsuka, Zhicheng Jing, and Mainak Mookherjee. "Shear deformation of dry polycrystalline olivine under deep upper mantle conditions using a rotational Drickamer apparatus (RDA)". In: *Phys. Earth. Planet. Inter.* 174.1-4 (2009), pp. 128–137. DOI: 10.1016/j.pepi.2008.06.027.

[727] N. Kechkar and D. Silvester. "Analysis of locally stabilised mixed finite element methods for the Stokes problem". In: *Mathematics of Computation* 58.197 (1992), pp. 1–10.

[728] T. Keller, D.A. May, and B.J.P. Kaus. "Numerical modelling of magma dynamics coupled to tectonic deformation of lithosphere and crust". In: *Geophy. J. Int.* 195.3 (2013), pp. 1406–1442. DOI: 10.1093/gji/ggt306.

[729] Tobias Keller and Paul J Tackley. "Towards self-consistent modeling of the martian dichotomy: The influence of one-ridge convection on crustal thickness distribution". In: *Icarus* 202.2 (2009), pp. 429–443. DOI: 10.1016/j.icarus.2009.03.029.

[730] L. H. Kellogg and S. D. King. "The effect of temperature dependent viscosity on the structure of new plumes in the mantle: Results of a finite element model in a spherical, axisymmetric shell". In: *Earth Planet. Sci. Lett.* 148.1-2 (1997), pp. 13–26. DOI: 10.1016/S0012-821X(97)00025-3.

[731] LH Kellogg and DL Turcotte. "Mixing and the distribution of heterogeneities in a chaotically convecting mantle". In: *J. Geophys. Res.: Solid Earth* 95.B1 (1990), pp. 421–432. DOI: 10.1029/JB095iB01p00421.

[732] Louise H Kellogg. "Chaotic mixing in the Earth's mantle". In: *Advances in geophysics*. Vol. 34. 1993, pp. 1–33. DOI: 10.1016/S0065-2687(08)60433-1.

[733] Louise H Kellogg, Bradford H Hager, and Rob D Van Der Hilst. "Compositional stratification in the deep mantle". In: *Science* 283.5409 (1999), pp. 1881–1884. DOI: 10.1126/science.283.5409.1881.

[734] Louise H Kellogg and Cheryl A Stewart. "Mixing by chaotic convection in an infinite Prandtl number fluid and implications for mantle convection". In: *Physics of Fluids A: Fluid Dynamics* 3.5 (1991), pp. 1374–1378. DOI: 10.1063/1.858067.

[735] B.L.N. Kennett. "On the density distribution within the Earth". In: *Geophy. J. Int.* 132.2 (1998), pp. 374–382. DOI: 10.1046/j.1365-246x.1998.00451.x.

[736] B.L.N. Kennett, E.R. Engdahl, and R. Buland. "Travel times for global earthquake location and phase association". In: *Geophy. J. Int.* 122 (1995), pp. 108–124. DOI: 10.1111/j.1365-246X.1991.tb06724.x.

[737] D.F. Keppie, C.A. Currie, and C. Warren. "Subduction erosion modes: comparing finite element numerical models with the geological record". In: *Earth Planet. Sci. Lett.* 287 (2009), pp. 241–254. DOI: 10.1016/j.epsl.2009.08.009.

[738] A.S. Khan and S. Huang. *Continuum theory of plasticity*. Wiley and sons, 1995.

[739] Amir Khan, Dongyang Huang, C Durán, Paolo Angelo Sossi, Domenico Giardini, and Motohiko Murakami. "Evidence for a liquid silicate layer atop the Martian core". In: *Nature* 622.7984 (2023), pp. 718–723. DOI: 10.1038/s41586-023-06586-4.

[740] Amir Khan et al. "A geophysical perspective on the bulk composition of Mars". In: *J. Geophys. Res.: Planets* 123.2 (2018), pp. 575–611. DOI: 10.1002/2017JE005371.

[741] W.S. Kiefer and B. Hager. "Geoid anomalies and dynamic topography from convection in cylindrical geometry: applications to mantle plumes on Earth and Venus". In: *Geophy. J. Int.* 108 (1992), pp. 198–214. DOI: 10.1111/j.1365-246X.1992.tb00850.x.

[742] S. King et al. "A community benchmark for 2D Cartesian compressible convection in the Earth's mantle". In: *Geophy. J. Int.* 180 (2010), pp. 73–87.

[743] S. D. King. "On topography and geoid from 2-D stagnant lid convection calculations". In: *Geochem. Geophys. Geosyst.* 10.3 (2009), n/a–n/a. DOI: 10.1029/2008GC002250.

[744] S.D. King and D.L. Anderson. "An alternative mechanism of flood basalt formation". In: *Earth Planet. Sci. Lett.* 136 (1995), pp. 269–279.

[745] S.D. King, D.J. Frost, and D.C. Rubie. "Why cold slabs stagnate in the transition zone". In: *Geology* 43.3 (2015), pp. 231–234. DOI: 10.1130/G36320.1.

[746] Scott D King. "Mantle convection, the asthenosphere, and Earth's thermal history". In: *Geological Society of America Special Papers* 514 (2015), SPE514–07. DOI: 10.1130/2015.2514(07).

[747] Scott D King. "Reconciling laboratory and observational models of mantle rheology in geodynamic modelling". In: *Journal of Geodynamics* 100 (2016), pp. 33–50. DOI: 10.1016/j.jog.2016.03.005.

[748] Scott D King and Guy Masters. "An inversion for radial viscosity structure using seismic tomography". In: *Geophys. Res. Lett.* 19.15 (1992), pp. 1551–1554. DOI: 10.1029/92GL01700.

[749] Á Király, Clinton P Conrad, and LN Hansen. "Evolving viscous anisotropy in the upper mantle and its geodynamic implications". In: *Geochem. Geophys. Geosyst.* 21 (2020), e2020GC009159. DOI: 10.1029/2020GC009159.

[750] S.H. Kirby and A.K. Kronenberg. "Rheology of the lithosphere: Selected topics". In: *Reviews of Geophysics* 25.6 (1987). DOI: 10.1029/RG025i006p01219.

[751] Matthias Kirchhart, Sven Gross, and Arnold Reusken. "Analysis of an XFEM discretization for Stokes interface problems". In: *SIAM Journal on Scientific Computing* 38.2 (2016), A1019–A1043. DOI: 10.1137/15M1011779.

[752] P. Kloucek, B. Li, and M. Luskin. "Analysis of a class of nonconforming finite elements for crystalline microstructures". In: *Mathematics of Computation* 65.215 (1996), pp. 1111–1135. DOI: xxxx.

[753] Brigitte Knapmeyer-Endrun et al. "Thickness and structure of the martian crust from InSight seismic data". In: *Science* 373.6553 (2021), pp. 438–443. DOI: 10.1126/science.abf8966.

[754] Erik A. Kneller, Markus Albertz, Garry D. Karner, and Christopher A. Johnson. "Testing inverse kinematic models of paleocrustal thickness in extensional systems with high-resolution forward thermo-mechanical models". In: *Geochem. Geophys. Geosyst.* (2013).

[755] Matthew G. Knepley. *Computational Science I.* Rice University - Department of Computational and Applied Mathematics, 2017.

[756] P. Knobloch. "On Korn's inequality for nonconforming Finite Elements". In: *Technische Mechanik* 20.3 (2000), pp. 205–214. DOI: xxxx.

[757] P. Knobloch and L. Tobiska. "Stabilisation methods of bubble type for the Q_1/Q_1 element applied to the incompressible Navier-Stokes equations". In: *Mathematical Modelling and Numerical Analysis* 34.1 (2000), pp. 85–107. DOI: 10.1051/m2an:2000132.

[758] Petr Knobloch. "On the definition of the SUPG parameter". In: *Electronic Transactions on Numerical Analysis* 32 (2008), pp. 76–89. DOI: xxxx.

[759] Dana A Knoll and David E Keyes. "Jacobian-free Newton–Krylov methods: a survey of approaches and applications". In: *J. Comp. Phys.* 193.2 (2004), pp. 357–397. DOI: 10.1016/j.jcp.2003.08.010.

[760] G.M. Kobelkov and M.A. Olshanskii. "Effective preconditioning of Uzawa type schemes for a generalized Stokes problem". In: *Numer. Math.* 86 (2000), pp. 443–470.

[761] D. E. Koglin Jr., S. R. Ghias, S. D. King, G. T. Jarvis, and J. P. Lowman. "Mantle convection with reversing mobile plates: A benchmark study". In: *Geochem. Geophys. Geosyst.* 6.9 (2005). DOI: 10.1029/2005GC000924.

[762] Nils Kohl, Marcus Mohr, Sebastian Eibl, and Ulrich Rüde. "A massively parallel Eulerian-Lagrangian method for advection-dominated transport in viscous fluids". In: *SIAM J. SCI. COMPUT.* 44.3 (2022), pp. C260–C285. DOI: 10.1137/21M1402510.

[763] Dimitri Komatitsch, Seiji Tsuboi, Jeroen Tromp, A Levander, and G Nolet. "The spectral-element method in seismology". In: *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION* 157 (2005), p. 205.

[764] L. Komzsik and P. Poschmann. "Iterative solution techniques for finite element applications". In: *Finite Elements in Analysis and Design* 14 (1993), pp. 373–379.

[765] Henk Kooi and Christopher Beaumont. "Escarpment evolution on high-elevation rifted margins: Insights derived from a surface processes model that combines diffusion, advection, and reaction". In: *J. Geophys. Res.: Solid Earth* 99.B6 (1994), pp. 12191–12209. DOI: 10.1029/94JB00047.

[766] U Kopitzke et al. "Finite element convection models: comparison of shallow and deep mantle convection, and temperatures in the mantle". In: *Journal of Geophysics* 46.1 (1979), pp. 97–121. DOI: xxx.

[767] Jun Korenaga. "Energetics of mantle convection and the fate of fossil heat". In: *Geophys. Res. Lett.* 30.8 (2003). DOI: 10.1029/2003GL016982.

[768] J.R. Koseff and R.L. Street. "The Lid-Driven Cavity Flow: A Synthesis of Qualitative and Quantitative Observations". In: *J. Fluids Eng* 106 (1984), pp. 390–398.

[769] D. Kourounis, A. Fuchs, and O. Schenk. "Towards the Next Generation of Multiperiod Optimal Power Flow Solvers". In: *IEEE Transactions on Power Systems* PP.99 (2018), pp. 1–10. DOI: 10.1109/TPWRS.2017.2789187.

[770] L.I.G. Kovasznay. "Laminar flow behind a two-dimensional grid". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 44.1 (1948), pp. 58–62. DOI: 10.1017/S0305004100023999.

[771] Peter Kovesi. "Good Colour Maps: How to Design Them". In: *CoRR* abs/1509.03700 (2015). arXiv: 1509.03700. URL: http://arxiv.org/abs/1509.03700.

[772] M. Krabbendam. "Sliding of temperate basal ice on a rough, hard bed: creep mechanisms, pressure melting, and implications for ice streaming". In: *The Cryosphere* 10 (2016), pp. 1915–1932. DOI: 10.5194/tc-10-1915-2016.

[773] Rolf Krahl and Eberhard Bänsch. "COMPUTATIONAL COMPARISON BETWEEN THE TAYLOR–HOOD AND THE CONFORMING CROUZEIX–RAVIART ELEMENT". In: *Proceedings of ALGORITMY*. 2005, pp. 369–379. DOI: xxxx.

[774] S.C. Kramer, D.R. Davies, and C.R. Wilson. "Analytical solutions for mantle flow in cylindrical and spherical shells". In: *Geosci. Model. Dev.* 14 (2021), pp. 1899–1919. DOI: 10.5194/gmd-14-1899-2021.

[775] Stephan C Kramer, Cian R Wilson, and D Rhodri Davies. "An implicit free surface algorithm for geodynamical simulations". In: *Phys. Earth. Planet. Inter.* 194 (2012), pp. 25–37. DOI: 10.1016/j.pepi.2012.01.001.

[776] M. Kronbichler, T. Heister, and W. Bangerth. "High accuracy mantle convection simulation through modern numerical methods". In: *Geophy. J. Int.* 191 (2012), pp. 12–29. DOI: 10.1111/j.1365-246X.2012.05609.x.

[777] Marcin Krotkiewski and Marcin Dabrowski. "Parallel symmetric sparse matrix–vector product on scalar multi-core CPUs". In: *Parallel Computing* 36.4 (2010), pp. 181–198. DOI: 10.1016/j.parco.2010.02.003.

[778] E.J. Kubatko, B.A. Yeager, and A.L. Maggi. "New computationally efficient quadrature formulas for triangular prism elements". In: *Computers and Fluids* 73 (2013), pp. 187–201. DOI: 10.1016/j.compfluid.2013.01.002.

[779] Hsaio-Lan Kuo. "Solution of the non-linear equations of cellular convection and heat transport". In: *Journal of Fluid Mechanics* 10.4 (1961), pp. 611–634. DOI: 10.1017/S0022112061000408.

[780] NJ Kusznir and MHP Bott. "Stress concentration in the upper lithosphere caused by underlying visco-elastic creep". In: *Tectonophysics* 43.3-4 (1977), pp. 247–256. DOI: 10.1016/0040-1951(77)90119-6.

[781] NJ Kusznir and RG Park. "The strength of intraplate lithosphere". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 224–235. DOI: 10.1016/0031-9201(84)90048-7.

[782] Oh-In Kwon and C. Park. "A new quadrilateral MINI element for Stokes equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 48 (2014), pp. 955–968. DOI: 10.1051/m2an/2013129.

[783] Olga Aleksandrovna Ladyzhenskaya. *The mathematical theory of viscous incompressible flow.* Vol. 2. Gordon and Breach New York, 1969.

[784] B.P. Lamichhane. "A quadrilateral 'mini' finite element for the Stokes problem using a single bubble function". In: *International Journal of Numerical Analysis and Modeling* 14.6 (2017), pp. 869–878.

[785] Bishnu P Lamichhane. "A mixed finite element method for nearly incompressible elasticity and Stokes equations using primal and dual meshes with quadrilateral and hexahedral grids". In: *Journal of computational and applied mathematics* 260 (2014), pp. 356–363. DOI: 10.1016/j.cam.2013.09.056.

[786] Bishnu P Lamichhane. "Inf–sup stable finite-element pairs based on dual meshes and bases for nearly incompressible elasticity". In: *IMA Journal of Numerical Analysis* 29.2 (2009), pp. 404–420. DOI: 10.1093/imanum/drn013.

[787] W. Landry, L. Hodkinson, and S. Kientz. *Gale User Manual.* Tech. rep. CIG, VPAC, 2011.

[788] Ulrich Langer and Martin Neumüller. "Direct and iterative solvers". In: *Computational Acoustics.* Springer, 2018, pp. 205–251. DOI: 10.1007/978-3-319-59038-7_5.

[789] Hans Petter Langtangen, Timothy J Barth, and Michael Griebel. *Python scripting for computational science.* Vol. 3. Springer, 2008.

[790] BE Larock and LR Herrmann. "Improved Flux Prediction Using Low Order Finite Elements". In: *International Conference on Finite Elements in Water Resources, Part II.* 1976. DOI: xxxx.

[791] T.B. Larsen and D.A. Yuen. "Fast plumeheads: Temperature-dependent versus non-Newtonian rheology". In: *Geophys. Res. Lett.* 24.16 (1997), pp. 1995–1998. DOI: 10.1029/97GL01886.

[792] T.B. Larsen and D.A. Yuen. "Ultrafast upwelling bursting through the upper mantle". In: *Earth Planet. Sci. Lett.* 146.3-4 (1997), pp. 393–399. DOI: 10.1016/S0012-821X(96)00247-6.

[793] T.B. Larsen, D.A. Yuen, A.V. Malevsky, and J.L. Smedsmo. "Dynamics of strongly time-dependent convection with non-Newtonian temperature-dependent viscosity". In: *Phys. Earth. Planet. Inter.* 94.1-2 (1996), pp. 75–103. DOI: 10.1016/0031-9201(95)03082-4.

[794] Tine B Larsen, David A Yuen, and Andrei V Malevsky. "Dynamical consequences on fast subducting slabs from a self-regulating mechanism due to viscous heating in variable viscosity convection". In: *Geophys. Res. Lett.* 22.10 (1995), pp. 1277–1280. DOI: 10.1029/95GL01112.

[795] Tine B Larsen, David A Yuen, Jiří Moser, and Bengt Fornberg. "A high-order finite-difference method applied to large Rayleigh number mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 84.1-2 (1997), pp. 53–83. DOI: 10.1080/03091929708208973.

[796] Mats G Larson and Fredrik Bengzon. *The finite element method: theory, implementation, and applications – Texts in Computational Science and Engineering.* Vol. 10. Springer Science & Business Media, 2013.

[797] Gautier Laurent, Guillaume Caumon, and Mark Jessell. "Interactive editing of 3D geological structures and tectonic history sketching via a rigid element method". In: *Computers and Geosciences* 74 (2015), pp. 71–86. DOI: 10.1016/j.cageo.2014.10.011.

[798] A. Lavecchia, C. Thieulot, F. Beekman, S. Cloetingh, and S. Clark. "Lithosphere erosion and continental breakup: Interaction of extension, plume upwelling and melting". In: *Earth Planet. Sci. Lett.* 467 (2017), pp. 89–98. DOI: 10.1016/j.epsl.2017.03.028.

[799] Luc L Lavier, W Roger Buck, and Alexei NB Poliakov. "Factors controlling normal fault offset in an ideal brittle layer". In: *J. Geophys. Res.: Solid Earth* 105.B10 (2000), pp. 23431–23442. DOI: 10.1029/2000JB900108.

[800] W. Layton. "Weak imposition of 'no-slip' conditions in finite element methods". In: *Computers and Mathematics with Applications* 38 (1999), pp. 129–142. DOI: 10.1016/S0898-1221(99)00220-5.

[801] Laetitia Le Pourhiet, Dave A May, Lucas Huille, Louise Watremez, and Sylvie Leroy. "A genetic link between transform and hyper-extended margins". In: *Earth Planet. Sci. Lett.* 465 (2017), pp. 184–192.

[802] E-S Lee, Charles Moulinec, Rui Xu, Damien Violeau, Dominique Laurence, and Peter Stansby. "Comparisons of weakly compressible and truly incompressible algorithms for the SPH mesh free particle method". In: *J. Comp. Phys.* 227.18 (2008), pp. 8417–8436. DOI: 10.1016/j.jcp.2008.06.005.

[803] Junhwan Lee, Rodrigo Salgado, and Sooil Kim. "Bearing capacity of circular footings under surcharge using state-dependent finite element analysis". In: *Computers and Geotechnics* 32 (2005), pp. 445–457. DOI: 10.1016/j.compgeo.2005.07.005.

[804] R. Lee, P. Gresho, and R. Sani. "Smoothing techniques for certain primitive variable solutions of the Navier-Stokes equations". In: *Int. J. Num. Meth. Eng.* 14 (1979), pp. 1785–1804.

[805] Seung Hyun Lee and Byung-Dal So. "Two dimensional finite element numerical model for slab detachment using Arbitrary Lagrangian Eulerian and remeshing". In: *J. Geol. Soc. Korea* 55.6 (2019), pp. 663–682. DOI: 10.14770/jgsk.2019.55.6.663.

[806] Young-Ju Lee and Hengguang Li. "Axisymmetric Stokes equations in polygonal domains: regularity and finite element approximations". In: *Computers & Mathematics with Applications* 64.11 (2012), pp. 3500–3521. DOI: 10.1016/j.camwa.2012.08.014.

[807] R.S. Lehmann, M. Lukacova-Medvidova, B.J.P. Kaus, and A.A. Popov. "Comparison of continuous and discontinuous Galerkin approaches for variable-viscosity Stokes flow". In: *Z. Angew. Math. Mech.* 96.6 (2015), pp. 733–746. DOI: 10.1002/zamm.201400274.

[808] A.M. Leitch, V. Steinbach, and D.A. Yuen. "Centerline temperature of mantle plumes in various geometries: Incompressible flow". In: *J. Geophys. Res.: Solid Earth* 101.B10 (1996), pp. 21829–21846. DOI: 10.1029/96JB01784.

[809] A.M. Leitch, D.A. Yuen, and G. Sewell. "Mantle convection with internal heating and pressure-dependent thermal expansivity". In: *Earth Planet. Sci. Lett.* 102.2 (1991), pp. 213–232. DOI: 10.1016/0012-821X(91)90009-7.

[810] V. Lemiale, H.-B. Mühlhaus, L. Moresi, and J. Stafford. "Shear banding analysis of plastic models formulated for incompressible viscous flows". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 177–186. DOI: 10.1016/j.pepi.2008.07.038.

[811] Vincent Lemiale, H-B Mühlhaus, Catherine Meriaux, L Moresi, and L Hodkinson. "Rate effects in dense granular materials: Linear stability analysis and the fall of granular columns". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 35.2 (2011), pp. 293–308. DOI: 10.1002/nag.895.

[812] A. Lenardic and W. M. Kaula. "A numerical treatment of geodynamic viscous flow problems involving the advection of material interfaces". In: *J. Geophys. Res.: Solid Earth* 98.B5 (1993), pp. 8243–8260. DOI: 10.1029/92JB02858.

[813] A. Lenardic and W. M. Kaula. "Near-surface thermal/chemical boundary layer convection at infinite Prandtl number: two-dimensional numerical experiments". In: *Geophy. J. Int.* 126.3 (1996), pp. 689–711. DOI: 10.1111/j.1365-246X.1996.tb04698.x.

[814] Adrian Lenardic, CM Cooper, and L Moresi. "A note on continents and the Earth's Urey ratio". In: *Phys. Earth. Planet. Inter.* 188.1-2 (2011), pp. 127–130. DOI: 10.1016/j.pepi.2011.06.008.

[815] W. Leng and S. Zhong. "Implementation and application of adaptive mesh refinement for thermochemical mantle convection studies". In: *Geochem. Geophys. Geosyst.* 12.4 (2011). DOI: 10.1029/2010GC003425.

[816] W. Leng and S. Zhong. "Viscous heating, adiabatic heating and energetic consistency in compressible mantle convection". In: *Geophy. J. Int.* 173 (2008), pp. 693–702. DOI: 10.1111/j.1365-246X.2008.03745.x.

[817] Wei Leng, Lili Ju, Yan Xie, Tao Cui, and Max Gunzburger. "Finite element three-dimensional Stokes ice sheet dynamics model with enhanced local mass conservation". In: *J. Comp. Phys.* 274 (2014), pp. 299–311.

[818] Y Leroy and M Ortiz. "Finite element analysis of strain localization in frictional materials". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 13.1 (1989), pp. 53–74. DOI: 10.1002/nag.1610130106.

[819] P. LeTallec. "Compatibility condition and existence results in discrete finite incompressible elasticity". In: *Computer Methods in Applied Mechanics and Engineering* 27 (1981), pp. 239–259.

[820] P. LeTallec and V. Ruas. "On the convergence of the bilinear-velocity constant-pressure finite element method in viscous flow". In: *Computer Methods in Applied Mechanics and Engineering* 54 (1986), pp. 235–243.

[821] Frank G Lether. "Computation of double integrals over a triangle". In: *Journal of Computational and Applied Mathematics* 2.3 (1976), pp. 219–224. DOI: 10.1016/0771-050X(76)90008-5.

[822] Einat Lev and Bradford H Hager. "Rayleigh–Taylor instabilities with anisotropic lithospheric viscosity". In: *Geophy. J. Int.* 173.3 (2008), pp. 806–814. DOI: 10.1111/j.1365-246X.2008.03731.x.

[823] R.J. Leveque. "High-resolution conservative algorithms for advection in incompressible flow". In: *SIAM J. Numer. Anal.* 33(2) (1996), pp. 627–665. DOI: 10.1137/0733033.

[824] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007. ISBN: 978-0-898716-29-0.

[825] Ben Q Li. *Discontinuous finite elements in fluid dynamics and heat transfer*. Springer Science & Business Media, 2006.

[826] D. Li, M. Gurnis, and G. Stadler. "Towards adjoint-based inversion of time-dependent mantle convection with nonlinear viscosity". In: *Geophy. J. Int.* 209 (2017), pp. 86–105. DOI: 10.1093/gji/ggw493.

[827] J. Li, Y. He, and Z. Chen. "Performance of several stabilized finite element methods for the Stokes equations based on the lowest equal-order pairs". In: *Computing* 86 (2009), pp. 37–51. DOI: 10.1007/s00607-009-0064-5.

[828] L. Li et al. "Deformation of olivine at mantle pressure using the D-DIA". In: *Eur. J. Mineral.* 18 (2006), pp. 7–19. DOI: 10.1127/0935-1221/2006/0018-0007.

[829] S. Li and W.K. Liu. *Meshfree Particle Methods*. Springer, 2004.

[830] Zhenyu Li and Roger E Khayat. "Finite-amplitude Rayleigh–Bénard convection and pattern selection for viscoelastic fluids". In: *Journal of Fluid Mechanics* 529 (2005), pp. 221–251. DOI: 10.1017/S0022112005003563.

[831] Zhenyu Li and Roger E Khayat. "Three-dimensional thermal convection of viscoelastic fluids". In: *Physical Review E* 71.6 (2005), p. 066305. DOI: 10.1103/PhysRevE.71.066305.

[832] Jie Liao and Taras Gerya. "Partitioning of crustal shortening during continental collision: 2-D thermomechanical modeling". In: *J. Geophys. Res.: Solid Earth* 122.1 (2017), pp. 592–606. DOI: 10.1002/2016JB013398.

[833] Q. Liao and D. Silvester. "Robust stabilized Stokes approximation methods for highly stretched grids". In: *IMA Journal of Numerical Analysis* 33 (2013), pp. 413–431. DOI: 10.1093/imanum/drs012.

[834] A. Limache, S. Idelsohn, R. Rossi, and E. Oñate. "The violation of objectivity in Laplace formulations of the Navier-Stokes equations". In: *Int. J. Num. Meth. Fluids* 54 (2007), pp. 639–664. DOI: 0.1002/fld.1480.

[835] S.-C. Lin and P.E. van Keken. "Dynamics of thermochemical plumes: 1. Plume formation and entrainment of a dense layer". In: *Geochem. Geophys. Geosyst.* 7.2 (2006). DOI: 10.1029/2005GC001071.

[836] S.-C. Lin and P.E. van Keken. "Dynamics of thermochemical plumes: 2. Complexity of plume structures and its implications for mapping mantle plumes". In: *Geochem. Geophys. Geosyst.* 7.3 (2006). DOI: 10.1029/2005GC001072.

[837] Shu-Chuan Lin and Peter E van Keken. "Multiple volcanic episodes of flood basalts caused by thermochemical mantle plumes". In: *Nature* 436.7048 (2005), pp. 250–252. DOI: 10.1038/nature03697.

[838] Johannes Linden, Guy Lonsdale, Barbara Steckel, and Klaus Stueben. "Multigrid for the steady-state incompressible navier-stokes equations: A survey". In: *11th International Conference on Numerical Methods in Fluid Dynamics*. Springer. 1989, pp. 57–68. DOI: xxxx.

[839] Konstantin D Litasov and Eiji Ohtani. "Effect of water on the phase relations in Earth's mantle and deep water cycle". In: *Special Papers-Geological Society of America* 421 (2007), p. 115. DOI: 10.1130/2007.2421(08).

[840] Benjamin T Liu, Susan J Muller, and Morton M Denn. "Convergence of a regularization method for creeping flow of a Bingham material about a rigid sphere". In: *Journal of non-newtonian fluid mechanics* 102.2 (2002), pp. 179–191. DOI: 10.1016/S0377-0257(01)00177-X.

[841] Chengjun Liu, Chengli Huang, and Mian Zhang. "The principal moments of inertia calculated with the hydrostatic equilibrium figure of the Earth". In: *Geodesy and Geodynamics* 8.3 (2017), pp. 201–205. DOI: 10.1016/j.geog.2017.02.005.

[842] G.R. Liu. *Mesh Free Methods*. CRC press, 2003.

[843] G.R. Liu and Y.T. Gu. *An introduction to meshfree methods and their programming*. Springer, 2005.

[844] G.R. Liu and M.B. Liu. *Smoothed Particle Hydrodynamics*. World Scientific, 2003.

[845] GR Liu and ZH Tu. "An adaptive procedure based on background cells for meshless methods". In: *Computer Methods in Applied Mechanics and Engineering* 191.17-18 (2002), pp. 1923–1943. DOI: 10.1016/S0045-7825(01)00360-7.

[846] P. Ho-Liu, B.H. Hager, and A. Raefsky. "An improved method of Nusselt number calculation". In: *Geophys. J. R. astr. Soc.* 88 (1987), pp. 205–215. DOI: `10.1111/j.1365-246X.1987.tb01375.x`.

[847] S. Liu and S.D. King. "A benchmark study of incompressible Stokes flow in a 3-D spherical shell using ASPECT". In: *Geophy. J. Int.* 217 (2019), pp. 650–667. DOI: `10.1093/gji/ggz036`.

[848] X. Liu and S. Zhong. "Analyses of marginal stability, heat transfer and boundary layer properties for thermal convection in a compressible fluid with infinite Prandtl number". In: *Geophy. J. Int.* 194 (2013), pp. 125–144. DOI: `10.1093/gji/ggt117`.

[849] Xi Liu and Shijie Zhong. "Constraining mantle viscosity structure for a thermochemical mantle using the geoid observation". In: *Geochem. Geophys. Geosyst.* 17.3 (2016), pp. 895–913. DOI: `10.1002/2015GC006161`.

[850] IS Lobanov, I Yu Popov, AI Popov, and TV Gerya. "Numerical approach to the Stokes problem with high contrasts in viscosity". In: *Applied Mathematics and Computation* 235 (2014), pp. 17–25. DOI: `10.1016/j.amc.2014.02.084`.

[851] J. Lof and A.H. van den Boogaard. "Adaptive return mapping algorithms for J_2 elasto-viscoplastic flow". In: *Int. J. Num. Meth. Eng.* 51 (2001), pp. 1283–1298. DOI: `10.1002/nme.203`.

[852] Anders Logg, Kent-Andre Mardal, and Garth Wells. *Automated solution of differential equations by the finite element method: The FEniCS book*. Vol. 84. Springer Science & Business Media, 2012.

[853] C. Loiselet et al. "Subducting slabs: Jellyfishes in the Earth's mantle". In: *Geochem. Geophys. Geosyst.* 11.8 (2010). DOI: `10.1029/2010GC003172`.

[854] Benjamin Loret and Jean H Prevost. "Dynamic strain localization in elasto-(visco-) plastic solids, Part 1. General formulation and one-dimensional examples". In: *Computer Methods in Applied Mechanics and Engineering* 83.3 (1990), pp. 247–273. DOI: `10.1016/0045-7825(90)90073-U`.

[855] Frank Losasso, Ronald Fedkiw, and Stanley Osher. "Spatially adaptive techniques for level set methods and incompressible flow". In: *Computers & Fluids* 35.10 (2006), pp. 995–1010. DOI: `10.1016/j.compfluid.2005.01.006`.

[856] Aurelie Louis–Napoleon, Thomas Bonometti, Muriel Gerbault, Roland Martin, and Olivier Vanderhaeghe. "Models of convection and segregation in heterogeneous partially molten crustal roots with a VOF method–I: flow regimes". In: *Geophy. J. Int.* 229.3 (2022), pp. 2047–2080. DOI: `10.1093/gji/ggab510`.

[857] Aurelie Louis–Napoleon, Muriel Gerbault, Thomas Bonometti, Cedric Thieulot, Roland Martin, and Olivier Vanderhaeghe. "3D numerical modeling of crustal polydiapirs with Volume-Of-Fluid methods". In: *Geophy. J. Int.* 222 (2020), pp. 474–506. DOI: `10.1093/gji/ggaa141`.

[858] Diogo L Lourenço, Antoine B Rozel, Taras Gerya, and Paul J Tackley. "Efficient cooling of rocky planets by intrusive magmatism". In: *Nature Geoscience* 11.5 (2018), p. 322.

[859] Julian P Lowman and Gary T Jarvis. "Continental collisions in wide aspect ratio and high Rayleigh number two-dimensional mantle convection models". In: *J. Geophys. Res.: Solid Earth* 101.B11 (1996), pp. 25485–25497. DOI: `10.1029/96JB02568`.

[860] Julian P Lowman and Gary T Jarvis. "Mantle convection flow reversals due to continental collisions". In: *Geophys. Res. Lett.* 20.19 (1993), pp. 2087–2090. DOI: `10.1029/93GL02047`.

[861] Julian P Lowman and Gary T Jarvis. "Mantle convection models of continental collision and breakup incorporating finite thickness plates". In: *Phys. Earth. Planet. Inter.* 88.1 (1995), pp. 53–68.

[862] Jacob Lubliner. *Plasticity theory*. Courier Corporation, 2008. ISBN: 978-0-486-46290-5.

[863] Richard A Lux, Geoffrey F Davies, and John H Thomas. "Moving lithospheric plates and mantle convection". In: *Geophy. J. Int.* 58.1 (1979), pp. 209–228.

[864] F. Machado, F. Zinani, and S. Frey. "Herschel-Bulkley Fluid Flows Through a Sudden Axisymmetric Expansion via Galerkin Least-Squares Methodology". In: ().

[865] P. Machetel and D.A. Yuen. "Penetrative convective flows induced by internal heating and mantle compressibility". In: *J. Geophys. Res.: Solid Earth* 94.B8 (1989), pp. 10, 609–10, 626. DOI: 10.1029/JB094iB08p10609.

[866] S.J. Mackwell, M. E. Zimmerman, and D. L. Kohlstedt. "High-temperature deformation of dry diabase with application to tectonics on Venus". In: *J. Geophys. Res.: Solid Earth* 103 (1998), pp. 975–984. DOI: 10.1029/97JB02671.

[867] Christopher W Macosko. "Rheology Principles". In: *Measurements and Applications* (1994).

[868] K Mader. "Das Newtonsche Raumpotential prismatischer Körper und seine Ableitungen bis zur dritten Ordnung, Osterr". In: *Z. Vermess. Sonderheft* 11 (1951).

[869] M. Maffione, C. Thieulot, D.J.J. van Hinsbergen, A. Morris, O. Plümper, and W. Spakman. "Dynamics of intraoceanic subduction initiation: 1. Oceanic detachment fault inversion and the formation of supra-subduction zone ophiolites". In: *Geochem. Geophys. Geosyst.* 16 (2015), pp. 1753–1770. DOI: 10.1002/2015GC005746.

[870] R. Mahmood, N. Kousar, M. Yaqub, and K. Jabeen. "Numerical Simulations of the Square Lid Driven Cavity Flow of Bingham Fluids Using Nonconforming Finite Elements Coupled with a Direct Solver". In: *Advances in Mathematical Physics* (2017).

[871] P. Maierová. "Evolution of the Bohemian Massif: Insights from numerical modeling". PhD thesis. Charles University in Prague, 2012.

[872] Detlev Majewski et al. "The operational global icosahedral–hexagonal gridpoint model GME: Description and high-resolution tests". In: *Monthly Weather Review* 130.2 (2002), pp. 319–338. DOI: 10.1175/1520-0493(2002)130<0319:TOGIHG>2.0.CO;2.

[873] A.V. Malevsky and D.A. Yuen. "Characteristics-based methods applied to infinite Prandtl number thermal convection in the hard turbulent regime". In: *Physics of Fluids A* 3.9 (1991), pp. 2105–2115. DOI: 10.1063/1.857893.

[874] A.V. Malevsky and D.A. Yuen. "Plume structures in the hard-turbulent regime of three-dimensional infinite Prandtl number convection". In: *Geophys. Res. Lett.* 20.5 (1993), pp. 383–386. DOI: 10.1029/93GL00293.

[875] A.V. Malevsky and D.A. Yuen. "Strongly chaotic non-newtonian mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 65.1-4 (1992), pp. 149–171. DOI: 10.1080/03091929208225244.

[876] Andrei V Malevsky, David A Yuen, and LM Weyer. "Viscosity and thermal fields associated with strongly chaotic non-Newtonian thermal convection". In: *Geophys. Res. Lett.* 19.2 (1992), pp. 127–130.

[877] Jakob M Maljaars, Chris N Richardson, and Nathan Sime. "LEoPart: A particle library for FEniCS". In: *Computers & Mathematics with Applications* 81 (2021), pp. 289–315. DOI: 10.1016/j.camwa.2020.04.023.

[878] D.S. Malkus and T.J.R. Hughes. "Mixed finite element methods - reduced and selective integration techniques: a unification of concepts". In: *Comput. Meth. Appl. Mech. Eng.* 15 (1978), pp. 63–81. DOI: `10.1016/0045-7825(78)90005-1`.

[879] L.E. Malvern. *Introduction to the mechanics of a continuous medium.* Prentice-Hall, Inc., 1969.

[880] N.S. Mancktelow. "Tectonic pressure: Theoretical concepts and modelled examples". In: *Lithos* 103 (2008), pp. 149–177. DOI: `10.1016/j.lithos.2007.09.013`.

[881] Michael Manga. "Mixing of heterogeneities in the mantle: Effect of viscosity differences". In: *Geophys. Res. Lett.* 23.4 (1996), pp. 403–406. DOI: `10.1029/96GL00242`.

[882] J.-P. Marcotte. "Methodes iteratives pour la resolution, par Elements Finis, du probleme de Stokes non lineaire". PhD thesis. Ecole Polytechnique de Montreal, 2000.

[883] Kent Andre Mardal, Xue-Cheng Tai, and Ragnar Winther. "A robust finite element method for Darcy–Stokes flow". In: *SIAM Journal on Numerical Analysis* 40.5 (2002), pp. 1605–1631. DOI: `10.1137/S0036142901383910`.

[884] FO Marques and YY Podladchikov. "A thin elastic core can control large-scale patterns of lithosphere shortening". In: *Earth Planet. Sci. Lett.* 277.1-2 (2009), pp. 80–85. DOI: `10.1016/j.epsl.2008.10.009`.

[885] Robert S Marshall, Juan C Heinrich, and OC Zienkiewicz. "Natural convection in a square enclosure by a finite-element, penalty function method using primitive fluid variables". In: *Numerical Heat Transfer, Part B: Fundamentals* 1.3 (1978), pp. 315–330. DOI: `10.1080/10407797809412168`.

[886] Zdeněk Martinec, Detlef Wolf, et al. "Explicit form of the propagator matrix for a multi-layered, incompressible viscoelastic sphere". In: (1998).

[887] J.G. Masek and C. Duncan. "Minimum-work mountain building". In: *J. Geophys. Res.: Solid Earth* 103.B1 (1998), pp. 907–917. DOI: `10.1029/97JB03213`.

[888] Paolo Massimi, Alfio Quarteroni, and G Scrofani. "An adaptive finite element method for modeling salt diapirism". In: *Mathematical Models and Methods in Applied Sciences* 16.04 (2006), pp. 587–614. DOI: `10.1142/S0218202506001273`.

[889] A. Massmeyer, E. Di Giuseppe, A. Davaille, T. Rolf, and P.J. Tackley. "Numerical simulation of thermal plumes in a Herschel-Bulkley fluid". In: *Journal of Non-Newtonian Rheology* 195 (2013), pp. 32–45.

[890] Takeshi Matsumoto and Yoshibumi Tomoda. "Numerical simulation of the initiation of subduction at the fracture zone". In: *Journal of Physics of the Earth* 31.3 (1983), pp. 183–194. DOI: `10.4294/jpe1952.31.183`.

[891] Ctirad Matyska and David A Yuen. "Lower-mantle material properties and convection models of multiscale plumes". In: *Special Papers – Geological Society of America* 430 (2007), p. 137.

[892] Ctirad Matyska and David A Yuen. "Profiles of the Bullen parameter from mantle convection modelling". In: *Earth Planet. Sci. Lett.* 178.1-2 (2000), pp. 39–46. DOI: `10.1016/S0012-821X(00)00060-1`.

[893] Ctirad Matyska, David A Yuen, Renata M Wentzcovitch, and Hana Čížková. "The impact of variability in the rheological activation parameters on lower-mantle viscosity stratification and its dynamics". In: *Phys. Earth. Planet. Inter.* 188.1-2 (2011), pp. 1–8. DOI: `10.1016/j.pepi.2011.05.012`.

[894] D.A. May, J. Brown, and L. Le Pourhiet. "A scalable, matrix-free multigrid preconditioner for finite element discretizations of heterogeneous Stokes flow". In: *Computer Methods in Applied Mechanics and Engineering* 290 (2015), pp. 496–523. DOI: 10.1016/j.cma.2015.03.014.

[895] D.A. May and L. Moresi. "Preconditioned iterative methods for Stokes flow problems arising in computational geodynamics". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 33–47. DOI: 10.1016/j.pepi.2008.07.036.

[896] D.A. May, W.P. Schellart, and L. Moresi. "Overview of adaptive finite element analysis in computational geodynamics". In: *Journal of Geodynamics* 70 (2013), pp. 1–20. DOI: 10.1016/j.jog.2013.04.002.

[897] Dave A May, Jed Brown, and Laetitia Le Pourhiet. "pTatin3D: High-performance methods for long-term lithospheric dynamics". In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*. IEEE Press. 2014, pp. 274–284. DOI: 10.1109/SC.2014.28.

[898] Neil McBride and Iain Gilmour. *The Solar System: Part 2*. Open University, 2003.

[899] J.L. McGregor. "Semi-Lagrangian advection of conformal-cubic grids". In: *Monthly Weather Review* 124 (1996), pp. 1311–1322.

[900] S. McKee et al. "The MAC method". In: *Computers and Fluids* 37 (2008), pp. 907–930. DOI: 10.1016/j.compuid.2007.10.006.

[901] D.P. McKenzie. "Speculations on the consequences and causes of plate motions". In: *Geophys. J. R. astr. Soc.* 18 (1969), pp. 1–32. DOI: 10.1111/j.1365-246X.1969.tb00259.x.

[902] Dan McKenzie. "Finite deformation during fluid flow". In: *Geophy. J. Int.* 58.3 (1979), pp. 689–715. DOI: 10.1111/j.1365-246X.1979.tb04803.x.

[903] DAN Mckenzie and MJ Bickle. "The volume and composition of melt generated by extension of the lithosphere". In: *Journal of petrology* 29.3 (1988), pp. 625–679. DOI: 10.1093/petrology/29.3.625.

[904] Dan McKenzie, Jean Roberts, and Nigel Weiss. "Numerical models of convection in the earth's mantle". In: *Tectonophysics* 19.2 (1973), pp. 89–103. DOI: 10.1016/0040-1951(73)90034-6.

[905] Dan P McKenzie. "Some remarks on heat flow and gravity anomalies". In: *J. Geophys. Res.: Solid Earth* 72.24 (1967), pp. 6261–6273. DOI: 10.1029/JZ072i024p06261.

[906] Dan P McKenzie, Jean M Roberts, and Nigel O Weiss. "Convection in the Earth's mantle: towards a numerical simulation". In: *Journal of Fluid Mechanics* 62.3 (1974), pp. 465–538. DOI: 10.1017/S0022112074000784.

[907] Colin P McNally. "Divergence-free interpolation of vector fields from point values - exact $\nabla \cdot B = 0$ in numerical simulations". In: *Monthly Notices of the Royal Astronomical Society: Letters* 413.1 (2011), pp. L76–L80.

[908] A. K. McNamara and S. Zhong. "Thermochemical structures within a spherical mantle: Superplumes or piles?" In: *J. Geophys. Res.: Solid Earth* 109.B7 (2004). DOI: 10.1029/2003JB002847.

[909] S. Mei, A.M. Suzuki, D.L. Kohlstedt, N.A. Dixon, and W.B. Durham. "Experimental constraints on the strength of the lithospheric mantle". In: *J. Geophys. Res.: Solid Earth* 115.B08204 (2010). DOI: 10.1029/2009JB006873.

[910] Mark van der Meijde, Roland Pail, R Bingham, and R Floberghagen. "GOCE data, models, and applications: A review". In: *International journal of applied earth observation and geoinformation* 35 (2015), pp. 4–15. DOI: 10.1016/j.jag.2013.10.001.

[911] HJ Melosh and Arthur Raefsky. "The dynamical origin of subduction zone topography". In: *Geophy. J. Int.* 60.3 (1980), pp. 333–354.

[912] HJ Melosh and CA Williams Jr. "Mechanics of graben formation in crustal rocks: A finite element analysis". In: *J. Geophys. Res.: Solid Earth* 94.B10 (1989), pp. 13961–13973.

[913] C.A. Mériaux, A. May D, J. Mansour, Z. Chen, and O. Kaluza. "Benchmark of three-dimensional numerical models of subduction against a laboratory experiment". In: *Phys. Earth. Planet. Inter.* 283 (2018), pp. 110–121.

[914] Guy Metcalfe, Craig R Bina, and JM Ottino. "Kinematic considerations for mantle mixing". In: *Geophys. Res. Lett.* 22.7 (1995), pp. 743–746. DOI: 10.1029/95GL00056.

[915] AJF Metherell and TJ Quinn. "The gravitational field of a 111 tetrahedron". In: *Metrologia* 22.2 (1986), p. 87. DOI: 10.1088/0026-1394/22/2/003.

[916] M.J. van der Meulen, S.J.H. Buiter, J.E. Meulenkamp, and M.J.R. Wortel. "An early Pliocene uplift of the central Apenninic foredeep and its geodynamic significance". In: *Tectonics* 19 (2000), pp. 300–313. DOI: 10.1029/1999TC900064.

[917] D.W. Meyer and P. Jenny. "Conservative Velocity Interpolation for PDF Methods". In: *Proc. Appl. Math. Mech.* 4 (2004), pp. 466–467.

[918] Victor Benno Meyer-Rochow and Jozsef Gal. "Pressures produced when penguins pooh - calculations on avian defaecation". In: *Polar Biology* 27.1 (2003), pp. 56–58.

[919] R. L. Michalowski. "Upper-bound load estimates on square and rectangular footings". In: *Géotechnique* 51.9 (2001), pp. 787–798.

[920] K. Michibayashi and D. Mainprice. "The role of pre-existing mechanical anisotropy on shear zone development within oceanic mantle lithosphere: an example from the Oman ophiolite". In: *J. Petrol.* 45(2) (2004), pp. 405–414.

[921] Laurent Michon and Olivier Merle. "Crustal structures of the Rhinegraben and the Massif Central grabens: An experimental approach". In: *Tectonics* 19.5 (2000), pp. 896–904. DOI: 10.1029/2000TC900015.

[922] Chohong Min and Frédéric Gibou. "A second order accurate level set method on non-graded adaptive cartesian grids". In: *J. Comp. Phys.* 225.1 (2007), pp. 300–321.

[923] P. Ming, Z-c. Shi, and Y. Xu. "A new superconvergence property of nonconforming rotated Q_1 element in 3D". In: *Computer Methods in Applied Mechanics and Engineering* 197 (2007), pp. 95–102.

[924] P. Ming, Z-c. Shi, and Y. Xu. "Superconvergence studies of quadrilateral nonconforming rotated Q_1 elements". In: *International Journal of Numerical analysis and modeling* 3.3 (2006), pp. 322–332.

[925] R. von Mises. "Mechanik der festen Korper im plastisch deformablen Zustand". In: *Nachrichten der Koniglichen Gesellschaft der Wissenchaften* (1913), p. 582.

[926] Y. Mishin. "Adaptive multiresolution methods for problems of computational geodynamics". PhD thesis. ETH Zurich, 2011.

[927] Y. Mishin. "Adaptive multiresolution methods for problems of computational geodynamics". PhD thesis. ETH Zürich, 2011. DOI: 10.3929/ethz-a-007347901.

[928] Y.A. Mishin, O.V. Vasilyev, and T.V. Gerya. "A Wavelet-Based Adaptive Finite Element Method for the Stokes Problems". In: *Fluids* 7 (2022), p. 221. DOI: 10.3390/fluids7070221.

[929] J.X. Mitrovica and A.M. Forte. "A new inference of mantle viscosity based upon joint inversion of convection and glacial isostatic adjustment data". In: *Earth Planet. Sci. Lett.* 225.1-2 (2004), pp. 177–189. DOI: 10.1016/j.epsl.2004.06.005.

[930] J.X. Mitrovica, R.N. Pysklywec, C. Beaumont, and A. Rutty. "The Devonian to Permian sedimentation of the Russian platform: An example of subduction-controlled long-wavelength tilting of continents". In: *Journal of Geodynamics* 22.1-2 (1996), pp. 79–96. DOI: 10.1016/0264-3707(96)00008-7.

[931] Jerry X Mitrovica. "Haskell [1935] revisited". In: *J. Geophys. Res.: Solid Earth* 101.B1 (1996), pp. 555–569. DOI: 10.1029/95JB03208.

[932] JX Mitrovica and GT Jarvis. "A numerical study of thermal convection between rigid horizontal boundaries". In: *Geophysical & Astrophysical Fluid Dynamics* 38.3 (1987), pp. 193–223. DOI: 10.1080/03091928708219204.

[933] JX Mitrovica and GT Jarvis. "Surface deflections due to transient subduction in a convecting mantle". In: *Tectonophysics* 120.3-4 (1985), pp. 211–237. DOI: 10.1016/0040-1951(85)90052-6.

[934] E. Mitsoulis and S. Galazoulas. "Simulation of viscoplastic flow past cylinders in tubes". In: *Journal of Non-Newtonian Fluid Mechanics* 158 (2009), pp. 132–141. DOI: 10.1016/j.jnnfm.2008.10.006.

[935] E. Mitsoulis and Th. Zisis. "Flow of Bingham plastics in a lid-driven square cavity". In: *Journal of Non-Newtonian Fluid Mechanics* 101 (2001), pp. 173–180.

[936] Arata Miyauchi and Masanori Kameyama. "Influences of the depth-dependence of thermal conductivity and expansivity on thermal convection with temperature-dependent viscosity". In: *Phys. Earth. Planet. Inter.* 223 (2013), pp. 86–95.

[937] A. Mizukami. "A mixed Finite Element method for boundary flux computation". In: *Computer Methods in Applied Mechanics and Engineering* 57 (1986), pp. 239–243.

[938] Arash Mohajeri, Yaron Finzi, Hans Mühlhaus, and Gideon Rosenbaum. "Melt and shear interactions in the lithosphere: Theory and numerical analysis of pure shear extension". In: *J. Geophys. Res.: Solid Earth* 118.5 (2013), pp. 2488–2499. DOI: 10.1111/j.1365-246X.2006.03225.x.

[939] P.K. Mohapatra, V. Eswaran, and S. Murty Bhallamudi. "Two-dimensional analysis of dam-break flow in vertical plane". In: *Journal of Hydraulic Engineering* 125.2 (1999), pp. 183–192.

[940] P. Molnar. *Brace-Goetze strength profiles, the partitioning of strike-slip and thrust faulting at zones of oblique convergence, and the stress-heat flow paradox of the San Andreas Fault.* Academic Press Ltd, 1992.

[941] P. Molnar and P. Tapponnier. "Relation of the tectonics of eastern China to the India-Eurasia collision: Application of the slip-line field theory to large-scale continental tectonics". In: *Geology* 5 (1977), pp. 212–216.

[942] JJ Monaghan. "Particle methods for hydrodynamics". In: *Computer Physics Reports* 3.2 (1985), pp. 71–124. DOI: 10.1016/0167-7977(85)90010-3.

[943] Puskar Mondal and Jun Korenaga. "A propagator matrix method for the Rayleigh–Taylor instability of multiple layers: a case study on crustal delamination in the early Earth". In: *Geophy. J. Int.* 212.3 (2018), pp. 1890–1901. DOI: 10.1093/gji/ggx513.

[944] P. Mons and G. Rogé. "L'élément Q_1-bulle/Q_1". In: *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* 26.4 (1992), pp. 507–521. DOI: xxxx.

[945] A. Montlaur, S. Fernandez-Mendez, and A. Huerta. "Discontinuous Galerkin methods for the Stokes equations using divergence-free approximations". In: *Int. J. Num. Meth. Fluids* 57.08 (2008), pp. 1071–1092.

[946] A. Montlaur, S. Fernandez-Mendez, J. Peraire, and A. Huerta. "Discontinuous Galerkin methods for the Navier-Stokes equations using solenoidal approximations". In: *Int. J. Num. Meth. Fluids* 64 (2010), pp. 549–564. DOI: 10.1002/fld.2161.

[947] Walter D Mooney, Gabi Laske, and T Guy Masters. "CRUST 5.1: A global crustal model at 5× 5". In: *J. Geophys. Res.: Solid Earth* 103.B1 (1998), pp. 727–747. DOI: 10.1029/97JB02122.

[948] L Moresi and H-B Mühlhaus. "Anisotropic viscous models of large-deformation Mohr–Coulomb failure". In: *Philosophical Magazine* 86.21-22 (2006), pp. 3287–3305. DOI: 10.1080/14786430500255419

[949] L Moresi, H-B Mühlhaus, Vincent Lemiale, and D May. "Incompressible viscous formulations for deformation and yielding of the lithosphere". In: *Geological Society, London, Special Publications* 282.1 (2007), pp. 457–472. DOI: 10.1144/SP282.19.

[950] L-N Moresi and A. Lenardic. "Three-dimensional numerical simulations of crustal deformation and subcontinental mantle convection". In: *Earth Planet. Sci. Lett.* 150.3-4 (1997), pp. 233–243. DOI: 10.1016/S0012-821X(97)00093-9.

[951] L. Moresi, F. Dufour, and H.B. Mühlhaus. "A Lagrangian integration point finite element method for large deformation modeling of visco-elastic geomaterials". In: *J. Comp. Phys.* 184.2 (2003), pp. 476–497. DOI: 10.1016/S0021-9991(02)00031-1.

[952] L. Moresi, F. Dufour, and H.B. Mühlhaus. "Mantle Convection Modeling with Viscoelastic/Brittle Lithosphere: Numerical Methodology and Plate Tectonic Modeling". In: *Pure Appl. Geophys.* 159 (2002), p. 159. DOI: 10.1007/s00024-002-8738-3.

[953] L. Moresi and M. Gurnis. "Constraints on the lateral strength of slabs from three-dimensional dynamic flow models". In: *Earth Planet. Sci. Lett.* 138.1 (1996), pp. 15–28. DOI: 10.1016/0012-821X(95)00221-W.

[954] L. Moresi, S. Quenette, V. Lemiale, C. Mériaux, B. Appelbe, and H.-B. Mühlhaus. "Computational approaches to studying non-linear dynamics of the crust and mantle". In: *Phys. Earth. Planet. Inter.* 163 (2007), pp. 69–82. DOI: 10.1016/j.pepi.2007.06.009.

[955] L. Moresi, S. Zhong, and M. Gurnis. "The accuracy of finite element solutions of Stokes' flow with strongly varying viscosity". In: *Phys. Earth. Planet. Inter.* 97.1-4 (1996), pp. 83–94. DOI: 10.1016/0031-9201(96)03163-9.

[956] L.-N. Moresi and V.S. Solomatov. "Numerical investigation of 2D convection with extremely large viscosity variations". In: *Physics of Fluids* 7.9 (1995), pp. 2154–2162.

[957] Louis Moresi and Ben Mather. "Stripy: A Python module for (constrained) triangulation in Cartesian coordinates and on a sphere". In: *Journal of Open Source Software* 4.38 (2019), p. 1410. DOI: 10.21105/joss.01410.

[958] Isabelle Moretti and Claude Froidevaux. "Thermomechanical models of active rifting". In: *Tectonics* 5.4 (1986), pp. 501–511.

[959] Jason P Morgan, Jorge M Taramón, and Jörg Hasenclever. "Shape-preserving finite elements in cylindrical and spherical geometries: The double Jacobian approach". In: *Int. J. Num. Meth. Fluids* 92.6 (2020), pp. 635–668. DOI: 10.1002/fld.4799.

[960] Gabriele Morra. "Pythonic Geodynamics". In: *Lecture Notes in Earth System Sciences* (2018).

[961] K.W. Morton and D.F. Mayers. *Numerical Solution of Partial Differential Equations: An Introduction*. Cambridge, 2005.

[962] Jiří Moser, Ctirad Matyska, David A Yuen, Andrei V Malevsky, and Helmut Harder. "Mantle rheology, convection and rotational dynamics". In: *Phys. Earth. Planet. Inter.* 79.3-4 (1993), pp. 367–381. DOI: 10.1016/0031-9201(93)90115-P.

[963] Robert Moucha et al. "Dynamic topography and long-term sea-level variations: There is no such thing as a stable continental platform". In: *Earth Planet. Sci. Lett.* 271.1-4 (2008), pp. 101–108. DOI: 10.1016/j.epsl.2008.03.056.

[964] P. Moulik and G. Ekström. "The relationships between large-scale variations in shear velocity, density, and compressional velocity in the Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 121 (2016). DOI: 10.1002/2015JB012679.

[965] Lin Mu and Xiu Ye. "A simple finite element method for the Stokes equations". In: *Advances in Computational Mathematics* 43.6 (2017), pp. 1305–1324. DOI: 10.1007/s10444-017-9526-z.

[966] H-B Mühlhaus, L Moresi, and Miroslav Cada. "Emergent anisotropy and flow alignment in viscous rock". In: *Pure Appl. Geophys.* 161.11-12 (2004), pp. 2451–2463.

[967] H-B Mühlhaus, L Moresi, and M Čada. "Anisotropy model for mantle convection". In: *Computational Fluid and Solid Mechanics 2003*. 2003, pp. 1044–1046. DOI: 10.1016/B978-008044046-0.50255-4.

[968] H.-B. Mühlhaus, L. Moresi, B. Hobbs, and F. Dufour. "Large amplitude folding in finely layered viscoelastic rock structures". In: *Pure Appl. Geophys.* 159 (2002), pp. 2311–2333.

[969] H.B. Mühlhaus and K. Regenauer-Lieb. "Towards a self-consistent plate mantle model that includes elasticity: simple benchmarks and application to basic modes of convection". In: *Geophy. J. Int.* 163 (2005), pp. 788–800. DOI: 10.1111/j.1365-246X.2005.02742.x.

[970] Hans Mühlhaus, Louis Moresi, Lutz Gross, and Joseph Grotowski. "The influence of non-coaxiality on shear banding in viscous-plastic materials". In: *Granular Matter* 12.3 (2010), pp. 229–238. DOI: 10.1007/s10035-010-0176-9.

[971] Hans B Mühlhaus, Jingyu Shi, Louise Olsen-Kettle, and Louis Moresi. "Effects of a non-coaxial flow rule on shear bands in viscous-plastic materials". In: *Granular Matter* 13.3 (2011), pp. 205–210.

[972] E. Mulyukova, B. Steinberger, M. Dabrowski, and S.V. Sobolev. "Survival of LLSVPs for billions of years in a vigorously convecting mantle: Replenishment and destruction of chemical anomaly". In: *J. Geophys. Res.: Solid Earth* 120 (2015), pp. 3824–3847. DOI: 10.1002/2014JB011688.

[973] Elvira Mulyukova. "Stability of the large low shear velocity provinces: numerical modeling of thermochemical mantle convection". PhD thesis. Universität Potsdam, 2014.

[974] K Nafa and RW Thatcher. "Low-order macroelements for two-and three-dimensional Stokes flow". In: *Numerical Methods for Partial Differential Equations* 9.5 (1993), pp. 579–591. DOI: 10.1002/num.1690090507.

[975] D. Nagy, G. Papp, and J. Benedek. "Corrections to "The gravitational potential and its derivatives for the prism"". In: *Journal of Geodesy* 76.8 (2002), p. 475. DOI: xxxx.

[976] D. Nagy, G. Papp, and J. Benedek. "The gravitational potential and its derivatives for the prism". In: *Journal of Geodesy* 74.7–8 (2000), pp. 552–560. DOI: 10.1007/s001900000116.

[977] Boris M Naimark, Ali T Ismail-Zadeh, and Wolfgang R Jacoby. "Numerical approach to problems of gravitational instability of geostructures with advected material boundaries". In: *Geophy. J. Int.* 134.2 (1998), pp. 473–483. DOI: 10.1111/j.1365-246X.1998.tb07140.x.

[978] Tomoeki Nakakuki and Erika Mura. "Dynamics of slab rollback and induced back-arc basin formation". In: *Earth Planet. Sci. Lett.* 361 (2013), pp. 287–297. DOI: 10.1016/j.epsl.2012.10.031.

[979] Tomoeki Nakakuki, Hiroki Sato, and Hiromi Fujimoto. "Interaction of the upwelling plume with the phase and chemical boundary at the 670 km discontinuity: Effects of temperature-dependent viscosity". In: *Earth Planet. Sci. Lett.* 121.3–4 (1994), pp. 369–384. DOI: 10.1016/0012-821X(94)90078-7.

[980] SM Nakiboglu and Kurt Lambeck. "A study of the Earth's response to surface loading with application to Lake Bonneville". In: *Geophy. J. Int.* 70.3 (1982), pp. 577–620. DOI: 10.1111/j.1365-246X.1982.tb05975.x.

[981] J. B. Naliboff and L. H. Kellogg. "Can large increases in viscosity and thermal conductivity preserve large-scale heterogeneity in the mantle?" In: *Phys. Earth. Planet. Inter.* 161.1-2 (2007), pp. 86–102. DOI: 10.1016/j.pepi.2007.01.009.

[982] J.B. Naliboff, S.J.H. Buiter, G. Péron-Pinvidic, P.T. Osmundsen, and J. Tetreault. "Complex fault interaction controls continental rifting". In: *Nature Communications* 8 (2017), p. 1179. DOI: 10.1038/s41467-017-00904-x.

[983] J.B. Naliboff, C. Lithgow-Bertelloni, L.J. Ruff, and N. de Koker. "The effects of lithospheric thickness and density structure on Earth's stress field". In: *Geophy. J. Int.* 188 (2012), pp. 1–17. DOI: 10.1111/j.1365-246X.2011.05248.x.

[984] Michele Napolitano, Giuseppe Pascazio, and Luigi Quartapelle. "A review of vorticity conditions in the numerical solution of the ζ–ψ equations". In: *Computers & Fluids* 28.2 (1999), pp. 139–185. DOI: 10.1016/S0045-7930(98)00024-3.

[985] A. Napov and Y. Notay. "An algebraic multigrid method with guaranteed convergence rate". In: *SIAM J. Sci. Comput.* 34 (2012), A1079–A1109. DOI: 10.1137/100818509.

[986] Mauricio Nava-Flores et al. "3D gravity modeling of complex salt features in the southern gulf of Mexico". In: *International Journal of Geophysics* 2016 (2016). DOI: 10.1155/2016/1702164.

[987] Jean-Claude Nédélec. "A new family of mixed finite elements in R^3". In: *Numerische Mathematik* 50 (1986), pp. 57–81. DOI: 10.1007/BF01389668.

[988] Jean-Claude Nédélec. "Mixed finite elements in R^3". In: *Numerische Mathematik* 35 (1980), pp. 315–341. DOI: 10.1007/BF01396415.

[989] A.M. Negredo, R. Sabadini, G. Bianco, and M. Fernandez. "Three-dimensional modelling of crustal motions caused by subduction and continental convergence in the central Mediterranean". In: *Geophy. J. Int.* 136 (1999), pp. 261–274. DOI: 10.1046/j.1365-246X.1999.00726.x.

[990] M. Nettesheim, T.A. Ehlers, D.M. Whipp, and A. Koptev. "The influence of upper-plate advance and erosion on overriding plate deformation in orogen syntaxes". In: *Solid Earth* 9 (2018), pp. 1207–1224. DOI: 10.5194/se-9-1207-2018.

[991] Derek Neuharth and Eric Mittelstaedt. "Temporal variations in plume flux: characterizing pulsations from tilted plume conduits in a rheologically complex mantle". In: *Geophy. J. Int.* 233.1 (2023), pp. 338–358. DOI: 10.1093/gji/ggac455.

[992] William I Newman. *Continuum Mechanics in the Earth Sciences*. Cambridge University Press, 2012. ISBN: 978-0-521-56289-8.

[993] N.C. Nguyen and J. Peraire. "Hybridizable discontinuous Galerkin methods for partial differential equations in continuum mechanics". In: *J. Comp. Phys.* 231 (2012), pp. 5955–5988. DOI: 10.1016/j.jcp.2012.02.033.

[994] N.C. Nguyen, J. Peraire, and B. Cockburn. "A hybridizable discontinuous Galerkin method for Stokes flow". In: *Computer Methods in Applied Mechanics and Engineering* 199 (2010), pp. 582–597. DOI: 10.1016/j.cma.2009.10.007.

[995]   N.C. Nguyen, J. Peraire, and B. Cockburn. "An implicit high-order hybridizable discontinuous Galerkin method for the incompressible Navier-Stokes equations". In: *J. Comp. Phys.* 230 (2011), pp. 1147–1170. DOI: 10.1016/j.jcp.2010.10.032.

[996]   G. Nicolas and T. Fouquet. "Adaptive mesh refinement for conformal hexahedral meshes". In: *Finite Elements in Analysis and Design* 67 (2013), pp. 1–12. DOI: 10.1016/j.finel.2012.11.008.

[997]   G. Nicolas and T. Fouquet. "Conformal hexahedral meshes and adaptive mesh refinement". In: *VI International Conference on Adaptive Modeling and Simulation* (2013). DOI: xxxx.

[998]   A.C. de Niet and F.W. Wubs. "Two preconditioners for the Saddle Point equation". In: *European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS 2004* (2004).

[999]   Nicolai Nijholt and Rob Govers. "The role of passive margins on the evolution of Subduction-Transform Edge Propagators (STEPs)". In: *J. Geophys. Res.: Solid Earth* 120.10 (2015), pp. 7203–7230. DOI: 10.1002/2015JB012202.

[1000] Lena Noack and Doris Breuer. "First-and second-order Frank-Kamenetskii approximation applied to temperature-, pressure-and stress-dependent rheology". In: *Geophy. J. Int.* 195.1 (2013), pp. 27–46. DOI: 10.1093/gji/ggt248.

[1001] S. Norburn and D. Silvester. "Fourier analysis of stabilized Q1-Q1 mixed finite element approximation". In: *SIAM J. Numer. Anal.* 39 (2001), pp. 817–833. DOI: 10.1137/S0036142999362274.

[1002] S. Norburn and D. Silvester. "Stabilised vs. stable mixed methods for incompressible flow". In: *Computer Methods in Applied Mechanics and Engineering* 166 (1998), pp. 131–141. DOI: 10.1016/S0045-7825(98)00087-5.

[1003] Y. Notay. "Aggregation-based algebraic multigrid for convection-diffusion equations". In: *SIAM J. Sci. Comput.* 34 (2012), A2288–A2316. DOI: 10.1137/110835347.

[1004] JL Nowinki. "On the Three-Dimensional Cerruti Problem for an Elastic Nonlocal Half-Space". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 72.7 (1992), pp. 243–249. DOI: 10.1002/zamm.19920720702.

[1005] K.A. O'Farrell, J.P. Lowman, and H.-P. Bunge. "Comparison of spherical-shell and plane-layer mantle convection thermal structure in viscously stratified models with mixed-mode heating: Implications for the incorporation of temperature-dependent parameters". In: *Geophy. J. Int.* 192.2 (2013), pp. 456–472. DOI: 10.1093/gji/ggs053.

[1006] C. O'Neill, L. Moresi, D. Müller, R. Albert, and F. Dufour. "Ellipsis 3D: a particle-in-cell finite element hybrid code for modelling mantle convection and lithospheric deformation". In: *Computers and Geosciences* 32 (2006), pp. 1769–1779. DOI: 10.1016/j.cageo.2006.04.006.

[1007] J Tinsley Oden and Olivier-Pierre Jacquotte. "Stability of some mixed finite element methods for Stokesian flows". In: *Computer Methods in Applied Mechanics and Engineering* 43.2 (1984), pp. 231–247.

[1008] J Tinsley Oden, Noboru Kikuchi, and Young Joon Song. "Penalty-finite element methods for the analysis of Stokesian flows". In: *Computer Methods in Applied Mechanics and Engineering* 31.3 (1982), pp. 297–329.

[1009] J. Oeser, H.-P. Bunge, M. Mohr, and H. Igel. "Frontiers in computational geophysics: Simulations of mantle circulation, plate tectonics and seismic wave propagation". In: *Notes on Numerical Fluid Mechanics and Multidisciplinary Design* 100 (2009), pp. 387–397. DOI: 10.1007/978-3-540-70805-6_30.

[1010] Masaki Ogawa. "A Numerical Model of a Coupled Magmatism-Mantle Convection System in Venus and the Earth's Mantle beneath Archean Continental Crusts". In: *Icarus* 102.1 (1993), pp. 40–61.

[1011] Masaki Ogawa, Gerald Schubert, and Abdelfattah Zebib. "Numerical simulations of three-dimensional thermal convection in a fluid with strongly temperature-dependent viscosity". In: *Journal of fluid mechanics* 233 (1991), pp. 299–328.

[1012] D. Olbertz, M.J.R. Wortel, and U. Hansen. "Trench migration and subduction zone geometry". In: *Geophys. Res. Lett.* 24.3 (1997), pp. 221–224. DOI: `10.1029/96GL03971`.

[1013] D. Oldham, J.H. Davies, and T.N. Phillips. "Generic polyhedron grid generation for solving partial differential equations on spherical surfaces". In: *Computers and Geosciences* 39 (2012), pp. 11–17. DOI: `10.1016/j.cageo.2011.06.004`.

[1014] Peter Olson. "A comparison of heat transfer laws for mantle convection at very high Rayleigh numbers". In: *Phys. Earth. Planet. Inter.* 48.1-2 (1987), pp. 153–160. DOI: `10.1016/0031-9201(87)90118-X`.

[1015] Peter Olson, Gerald Schubert, and Charles Anderson. "Structure of axisymmetric mantle plumes". In: *J. Geophys. Res.: Solid Earth* 98.B4 (1993), pp. 6829–6844. DOI: `10.1029/92JB01013`.

[1016] Peter Olson, David A Yuen, and Derick Balsiger. "Mixing of passive heterogeneities by mantle convection". In: *J. Geophys. Res.: Solid Earth* 89.B1 (1984), pp. 425–436. DOI: `10.1029/JB089iB01p00425`.

[1017] Peter Olson, David A. Yuen, and Derick Balsiger. "Convective mixing and the fine structure of mantle heterogeneity". In: *Phys. Earth. Planet. Inter.* 36.3–4 (1984), pp. 291–304. DOI: `10.1016/0031-9201(84)90053-0`.

[1018] E. Olsson and G. Kreiss. "A conservative level set method for two phase flow". In: *J. Comp. Phys.* 210 (2005), pp. 225–246.

[1019] E. Olsson, G. Kreiss, and S. Zahedi. "A conservative level set method for two phase flow II". In: *J. Comp. Phys.* 225 (2007), pp. 785–807.

[1020] S. Osher and R. Fedkiw. "Level set methods: an overview and some recent results". In: *J. Comp. Phys.* 169 (2001), pp. 463–502.

[1021] S. Osher and C.-W. Shu. "High-order essentially non-oscillatory schemes for Hamilton-Jacobi equations". In: *SIAM J. Numer. Anal* 28 (1991), pp. 907–922.

[1022] Tim A Osswald and Natalie Rudolph. *Polymer rheology: fundamentals and applications.* Carl Hanser Verlag GmbH Co KG, 2014.

[1023] J.M. Ottino. *The kinematics of mixing: stretching, chaos, and transport.* Vol. 3. Cambridge university press, 1989.

[1024] A. Ouazzi. "Finite Element Simulation of Nonlinear Fluids with Application to Granular Material and Powder". PhD thesis. Universität Dortmund, 2005.

[1025] Guillaume Ovarlez and Sarah Hormozi. *Lectures on visco-plastic fluid mechanics.* Springer, 2019. ISBN: 978-3-319-89437-9.

[1026] D.R.J. Owen and E. Hinton. *Finite Elements in Plasticity.* Pineridge press, 1980.

[1027] M. OzBench et al. "A model comparison study of large-scale mantle-lithosphere dynamics driven by subduction". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 224–234. DOI: `10.1016/j.pepi.2008.08.011`.

[1028] Karen Paczkowski, Laurent GJ Montési, Maureen D Long, and Christopher J Thissen. "Three-dimensional flow in the subslab mantle". In: *Geochem. Geophys. Geosyst.* 15.10 (2014), pp. 3989–4008. DOI: 10.1002/2014GC005441.

[1029] C.C. Paige and M.A. Saunders. "Solution of sparse indefinite systems of linear equations". In: *SIAM J. Numer. Anal* 12.4 (1975), pp. 617–629. DOI: 10.1137/0712047.

[1030] Isabelle Panet, Gwendoline Pajot-Métivier, Marianne Greff-Lefftz, Laurent Métivier, Michel Diament, and Mioara Mandea. "Mapping the mass distribution of Earth's mantle using satellite-derived gravity gradients". In: *Nature Geoscience* 7.2 (2014), pp. 131–135. DOI: 10.1038/NGEO2063.

[1031] Tasos C Papanastasiou. "Flows of materials with yield". In: *Journal of Rheology* 31.5 (1987), pp. 385–404.

[1032] Giovanni Pari and W Richard Peltier. "Global surface heat flux anomalies from seismic tomography-based models of mantle flow: Implications for mantle convection". In: *J. Geophys. Res.: Solid Earth* 103.B10 (1998), pp. 23743–23780. DOI: 10.1029/98JB01668.

[1033] Giovanni Pari and W Richard Peltier. "The heat flow constraint on mantle tomography-based convection models: Towards a geodynamically self-consistent inference of mantle viscosity". In: *J. Geophys. Res.: Solid Earth* 100.B7 (1995), pp. 12731–12751. DOI: 10.1029/95JB01078.

[1034] Chunjae Park and Dongwoo Sheen. "P_1-nonconforming quadrilateral finite element methods for second-order elliptic problems". In: *SIAM Journal on Numerical Analysis* 41.2 (2003), pp. 624–640. DOI: 10.1137/S0036142902404923.

[1035] Chunjae Park, Dongwoo Sheen, and Byeong-Chun Shin. "A subspace of the DSSY nonconforming quadrilateral finite element space for the Stokes equations". In: *Journal of Computational and Applied Mathematics* 239 (2013), pp. 220–230. DOI: 10.1016/j.cam.2012.09.042.

[1036] EM Parmentier, C Sotin, and BJ Travis. "Turbulent 3-D thermal convection in an infinite Prandtl number, volumetrically heated fluid: implications for mantle dynamics". In: *Geophy. J. Int.* 116.2 (1994), pp. 241–251. DOI: 10.1111/j.1365-246X.1994.tb01795.x.

[1037] EM Parmentier, DL Turcotte, and KE Torrance. "Numerical experiments on the structure of mantle plumes". In: *J. Geophys. Res.: Solid Earth* 80.32 (1975), pp. 4417–4424. DOI: 10.1029/JB080i032p04417.

[1038] EM Parmentier, DL Turcotte, and KE Torrance. "Studies of finite amplitude non-Newtonian thermal convection with application to convection in the Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 81.11 (1976), pp. 1839–1846. DOI: 10.1029/JB081i011p01839.

[1039] M.H. Parrish. "A SELECTIVE APPROACH TO CONFORMAL REFINEMENT OF UNSTRUCTURED HEXAHEDRAL MESHES". PhD thesis. Department of Civil and Environmental Engineering, Brigham Young University, 2007.

[1040] Barry Parsons and Stephen Daly. "The relationship between surface topography, gravity anomalies, and temperature structure of convection". In: *J. Geophys. Res.: Solid Earth* 88.B2 (1983), pp. 1129–1144.

[1041] S.V. Patankar and D.B. Spalding. "A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows". In: *Int. J. Heat Mass Transfer* 15 (1972), pp. 1787–1806. DOI: 10.1016/B978-0-08-030937-8.50013-1.

[1042] WSB Paterson and WF Budd. "Flow parameters for ice sheet modeling". In: *Cold Regions Science and Technology* 6.2 (1982), pp. 175–177. DOI: 10.1016/0165-232X(82)90010-6.

[1043] Archie Paulson, Shijie Zhong, and John Wahr. "Inference of mantle viscosity from GRACE and relative sea level data". In: *Geophy. J. Int.* 171.2 (2007), pp. 497–508. DOI: 10.1111/j.1365-246X.2007.03556.x.

[1044] Dorival M Pedroso. "Caveats of three direct linear solvers for finite element analyses". In: *Int. J. Num. Meth. Eng.* (2024), e7545. DOI: 10.1002/nme.7545.

[1045] D. Pelletier, A. Fortin, and R. Camarero. "Are FEM solutions of incompressible flows really incompressible ? (Or how simple flows can cause headaches!)" In: *Int. J. Num. Meth. Fluids* 9 (1989), pp. 99–112. DOI: 10.1002/fld.1650090108.

[1046] Jon Pelletier. *Quantitative modelling of Earth surface processes.* Cambridge University Press, 2008.

[1047] WR Peltier. "Penetrative convection in the planetary mantle". In: *Geophysical Fluid Dynamics* 5.1 (1972), pp. 47–88. DOI: 10.1080/03091927308236108.

[1048] WR Peltier. "The impulse response of a Maxwell Earth". In: *Reviews of Geophysics* 12.4 (1974), pp. 649–669. DOI: 10.1029/RG012i004p00649.

[1049] WR Peltier, S Butler, and LP Solheim. "The influence of phase transformations on mantle mixing and plate tectonics". In: *Earth's Deep Interior. Gordon & Breach, Amsterdam* (1997), pp. 405–430.

[1050] WR Peltier and GT Jarvis. "Whole mantle convection and the thermal evolution of the Earth". In: *Phys. Earth. Planet. Inter.* 29.3-4 (1982), pp. 281–304. DOI: 10.1016/0031-9201(82)90018-8.

[1051] G. Peltzer and P. Tapponnier. "Formation and evolution of strike-slip faults, rifts, and basins during the india-asia collision: an experimental approach". In: *J. Geophys. Res.: Solid Earth* 93.B12 (1988), pp. 15085–15177. DOI: 10.1029/JB093iB12p15085.

[1052] D. Perić and C. Huang. "Analytical solutions for the three-invariant cam clay model: drained and undrained loading". In: *16th ASCE Engineering Mechanics Conference* (2003).

[1053] J. Perry-Houts and L. Karlstrom. "Anisotropic viscosity and time-evolving lithospheric instabilities due to aligned igneous intrusions". In: *Geophy. J. Int.* 216.2 (2018), pp. 794–802. DOI: 10.1093/gji/ggy466.

[1054] Diego Perugini, CP De Campos, Maurizio Petrelli, Daniele Morgavi, Francesco Pasqualino Vetere, and DB Dingwell. "Quantifying magma mixing with the Shannon entropy: Application to simulations and experiments". In: *Lithos* 236 (2015), pp. 299–310. DOI: 10.1016/j.lithos.2015.09.008.

[1055] P. Perzyna. "Constitutive Modelling of Dissipative Solids for Localization and Fracture". In: *Localization and Fracture Phenomena in Inelastic Solids.* Ed. by Piotr Perzyna. Vienna: Springer Vienna, 1998, pp. 99–241. ISBN: 978-3-7091-2528-1. DOI: 10.1007/978-3-7091-2528-1_3.

[1056] Piotr Perzyna. "Fundamental problems in viscoplasticity". In: *Advances in applied mechanics.* Vol. 9. 1966, pp. 243–377. DOI: 10.1016/S0065-2156(08)70009-7.

[1057] A.G. Petrunin and S.V. Sobolev. "Three-dimensional numerical models of the evolution of pull-apart basins". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 387–399. DOI: 10.1016/j.pepi.2008.08.017.

[1058] Tim N Phillips, J Huw Davies, and D Oldham. "Towards global SEM mantle convection simulations on polyhedral-based grids". In: *Journal of Computational and Applied Mathematics* 348 (2019), pp. 48–57. DOI: 10.1016/j.cam.2018.08.042.

[1059] E. Pichelin and T. Coupez. "Finite element solution of the 3D mold filling problem for viscous incompressible fluid". In: *Computer Methods in Applied Mechanics and Engineering* 163 (1998), pp. 359–371. DOI: 10.1016/S0045-7825(98)00024-3.

[1060] RT Pierrehumbert. "Large-scale horizontal mixing in planetary atmospheres". In: *Physics of Fluids A: Fluid Dynamics* 3.5 (1991), pp. 1250–1260. DOI: 10.1063/1.858053.

[1061] D.A. di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods.* Springer, 2012.

[1062] D.A. Di Pietro, S. Lo Forte, and N. Parolini. "Mass preserving finite element implementations of the level set method". In: *Applied Numerical Mathematics* 56 (2006), pp. 1179–1195. DOI: 10.1016/j.apnum.2006.03.003.

[1063] J. Pitkäranta and T. Saarinen. "A Multigrid Version of a Simple Finite Element Method for the Stokes Problem". In: *Mathematics of Computation* 45.171 (1985), pp. 1–14.

[1064] A-C Plesa et al. "The thermal state and interior structure of Mars". In: *Geophys. Res. Lett.* 45.22 (2018), pp. 12–198. DOI: 10.1029/2018GL080728.

[1065] A.-C. Plesa. "Mantle Convection in a 2D Spherical Shell". In: *NFOCOMP 2011 : The First International Conference on Advanced Communications and Computation* (2011).

[1066] Ana-Catalina Plesa, Nicola Tosi, and Christian Hüttig. "Thermo-chemical convection in planetary mantles: advection methods and magma ocean overturn simulations". In: *Integrated Information and Computing Systems for Natural, Spatial, and Social Sciences.* IGI Global, 2013, pp. 302–323. DOI: 10.4018/978-1-4666-2190-9.ch015.

[1067] Ana-Catalina Plesa, Mark Wieczorek, Martin Knapmeyer, Attilio Rivoldini, Michaela Walterova, and Doris Breuer. "Interior dynamics and thermal evolution of Mars–a geodynamic perspective". In: *Advances in geophysics.* Vol. 63. 2022, pp. 179–230. DOI: 10.1016/bs.agph.2022.07.005.

[1068] A Plimmer, JH Davies, and J Panton. "Investigating the effect of lithosphere thickness and viscosity on mantle dynamics throughout the supercontinent cycle". In: *Geochem. Geophys. Geosyst.* 25.11 (2024), e2024GC011688. DOI: 10.1029/2024GC011688.

[1069] Donald Plouff. "Gravity and magnetic fields of polygonal prisms and application to magnetic terrain corrections". In: *Geophysics* 41.4 (1976), pp. 727–741. DOI: 10.1190/1.1440645.

[1070] Meredith Plumley and Keith Julien. "Scaling laws in Rayleigh-Benard convection". In: *Earth and Space Science* 6.9 (2019), pp. 1580–1592. DOI: 10.1029/2019EA000583.

[1071] Jean-Paul Poirier. *Introduction to the Physics of the Earth's Interior.* Cambridge University Press, 2000.

[1072] A. Poliakov, P. Cundall, P. Podlachikov, and V. Lyakhovsky. "An explicit inertial method for the simulation of viscoelastic flow: an evaluation of elastic effects on diapiric flow in two- and three-layers models". In: *Flow and creep in the solar system: Observations, Modeling and theory.* Kluwer Academic Publishers, 1993, pp. 175–195.

[1073] A. Poliakov and Y. Podlachikov. "Diapirism and topography". In: *Geophy. J. Int.* 109 (1992), pp. 553–564. DOI: 10.1111/j.1365-246X.1992.tb00117.x.

[1074] A.N.B. Poliakov and H.J. Herrmann. "Self-organized criticality of plastic shear bands in rocks". In: *Geophys. Res. Lett.* 21.19 (1994), pp. 2143–2146. DOI: 10.1029/94GL02005.

[1075] J-P Ponthot and Ted Belytschko. "Arbitrary Lagrangian-Eulerian formulation for element-free Galerkin method". In: *Computer Methods in Applied Mechanics and Engineering* 152.1-2 (1998), pp. 19–46. DOI: 10.1016/S0045-7825(97)00180-1.

[1076] A.A. Popov and S.V. Sobolev. "SLIM3D: a tool for three-dimensional thermomechanical modelling of lithospheric deformation with elasto-visco-plastic rheology". In: *Phys. Earth. Planet. Inter.* 171.1 (2008), pp. 55–75. DOI: 10.1016/j.pepi.2008.03.007.

[1077] AI Popov, IS Lobanov, I Yu Popov, and TV Gerya. "Benchmark solutions for nanoflows". In: *NANOSYSTEMS: PHYSICS, CHEMISTRY, MATHEMATICS* 5.3 (2014). DOI: xxxx.

[1078] I.Yu. Popov, I.S. Lobanov, S.I. Popov, A.I. Popov, and T. Gerya. "Practical analytical solutions for benchmarking of 2D and 3D geodynamic Stokes problems with variable viscosity". In: *Solid Earth* 5 (2014), pp. 461–476. DOI: 10.5194/se-5-461-2014.

[1079] Igor Yu Popov and Ilya V Makeev. "A benchmark solution for 2D Stokes flow over cavity". In: *Zeitschrift für angewandte Mathematik und Physik* 65.2 (2014), pp. 339–348.

[1080] Jean H Prevost and Benjamin Loret. "Dynamic strain localization in elasto-(visco-) plastic solids, part 2. Plane strain examples". In: *Computer Methods in Applied Mechanics and Engineering* 83.3 (1990), pp. 275–294. DOI: 10.1016/0045-7825(90)90074-V.

[1081] Matthew Price. "Investigating the initial condition of mantle models using data assimilation". PhD thesis. Cardiff University, 2016.

[1082] TE Price. "Numerically exact integration of a family of axisymmetric finite elements". In: *Communications in numerical methods in engineering* 19.4 (2003), pp. 253–261. DOI: 10.1002/cnm.583.

[1083] P.J. Prince and J.R. Dormand. "High order embedded Runge-Kutta formulae". In: *Journal of Computational and Applied Mathematics* 7.1 (1981), pp. 67–75. DOI: 10.1016/0771-050X(81)90010-3.

[1084] A. Prosperetti. "Motion of two superposed viscous fluids". In: *Phys. Fluids* 24.7 (1981), pp. 1217–1223.

[1085] E.G. Puckett, D.L. Turcotte, Y. He, H. Lokavarapu, J.M. Robey, and L.H. Kellogg. "New numerical approaches for modeling thermochemical convection in a compositionally stratified fluid". In: *Phys. Earth. Planet. Inter.* 276 (2018), pp. 10–35. DOI: 10.1016/j.pepi.2017.10.004.

[1086] Adina E Pusok, Boris JP Kaus, and Anton A Popov. "On the Quality of Velocity Interpolation Schemes for Marker-in-Cell Method and Staggered Grids". In: *Pure Appl. Geophys.* 174 (2017), pp. 1071–1089. DOI: 10.1007/s00024-016-1431-8.

[1087] W.M. Putman and S.-J. Lin. "Finite-Volume transport on various cubed-sphere grids". In: *J. Comp. Phys.* 227 (2007), pp. 55–78. DOI: 10.1016/j.jcp.2007.07.022.

[1088] R.N. Pysklywec, C. Beaumont, and P. Fullsack. "Lithospheric deformation during the early stages of continental collision: Numerical experiments and comparison with South Island, New Zealand". In: *J. Geophys. Res.: Solid Earth* 107.B72133 (2002). DOI: 10.1029/2001JB000252.

[1089] R.N. Pysklywec, C. Beaumont, and P. Fullsack. "Modeling the behavior of continental mantle lithosphere during plate convergence". In: *Geology* 28.7 (2000), pp. 655–658. DOI: 10.1130/0091-7613(2000)28<655:MTBOTC>2.0.CO;2.

[1090] Chuan Qin, Shijie Zhong, and Roger Phillips. "Formation of the lunar fossil bulges and its implication for the early Earth and Moon". In: *Geophys. Res. Lett.* 45.3 (2018), pp. 1286–1296. DOI: 10.1002/2017GL076278.

[1091] J. Qin and S. Zhang. "On the selective local stabilization of the mixed Q1-P0 element". In: *Int. J. Num. Meth. Eng.* 55.12 (2007), pp. 1121–1141. DOI: 10.1002/fld.1505.

[1092] J. Qin and S. Zhang. "Stability and approximability of the P1-P0 element for stokes equations". In: *Int. J. Num. Meth. Eng.* 54 (2007), pp. 497–515. DOI: 10.1002/fld.1407.

[1093] M.E.T. Quinquis and S.J.H. Buiter. "Testing the effects of basic numerical implementations of water migration on models of subduction dynamics". In: *Solid Earth* 5 (2014), pp. 537–555. DOI: 10.5194/se-5-537-2014.

[1094] Matthieu E.T. Quinquis, Suzanne J.H. Buiter, and Susan Ellis. "The role of boundary conditions in numerical models of subduction zone dynamics". In: *Tectonophysics* 497 (2011), pp. 57–70. DOI: 10.1016/j.tecto.2010.11.001.

[1095] J. Quinteros, V.A. Ramos, and P.M. Jacovkis. "An elasto-visco-plastic model using the finite element method for crustal and lithospheric deformation". In: *Journal of Geodynamics* 48 (2009), pp. 83–94. DOI: 10.1016/j.jog.2009.06.006.

[1096] J. Quinteros, S.V. Sobolev, and A.A. Popov. "Viscosity in transition zone and lower mantle: Implications for slab penetration". In: *Geophys. Res. Lett.* 37.L09307 (2010). DOI: 10.1029/2010GL043140.

[1097] T. Rabczuk, P.M.A. Areias, and T. Belytschko. "A simplified mesh-free method for shear bands with cohesive surfaces ". In: *Int. J. Num. Meth. Eng.* 69 (2007), pp. 993–1021. DOI: 10.1002/nme.1797.

[1098] M Rabinowicz, S Rouzo, J-C Sempere, and C Rosemberg. "Three-dimensional mantle flow beneath mid-ocean ridges". In: *J. Geophys. Res.: Solid Earth* 98.B5 (1993), pp. 7851–7869. DOI: 10.1029/92JB02740.

[1099] A. Ramachandran. "Parallel adaptive finite element simulation using distribued quad-tree/octree forest". MA thesis. Institute for Structural Mechanics, Ruhr University Bochum, 2016.

[1100] Balasubramaniam Ramaswamy. "Numerical simulation of unsteady viscous free surface flow". In: *J. Comp. Phys.* 90.2 (1990), pp. 396–430.

[1101] Balasubramaniam Ramaswamy and Mutsuto Kawahara. "Arbitrary Lagrangian–Eulerian finite element method for unsteady, convective, incompressible viscous free surface fluid flow". In: *Int. J. Num. Meth. Fluids* 7.10 (1987), pp. 1053–1075.

[1102] H. Ramberg. *Gravity, Deformation, and the Earth's Crust: In Theory, Experiments and Geological Application*. Academic Press, London, 214pp., 1967.

[1103] Hans Ramberg. "Folding of laterally compressed multilayers in the field of gravity, I". In: *Phys. Earth. Planet. Inter.* 2.4 (1970), pp. 203–232. DOI: 10.1016/0031-9201(70)90010-5.

[1104] Hans Ramberg. *Gravity, deformation and the earth's crust: in theory, experiments and geological application*. Academic press, 1981.

[1105] Hans Ramberg. "Instability of layered systems in the field of gravity". In: *Phys. Earth. Planet. Inter.* 1 (1968), pp. 427–447. DOI: 10.1016/0031-9201(68)90014-9.

[1106] Ekkehard Ramm et al. *Error-controlled adaptive finite elements in solid mechanics*. 2003.

[1107] G. Ranalli. "Rheology of the crust and its role in tectonic reactivation". In: *Journal of Geodynamics* 30 (2000), pp. 3–15.

[1108] G. Ranalli. *Rheology of the Earth*. Springer, 1995. ISBN: 0-412-54670-1.

[1109] G. Ranalli. "Rheology of the lithosphere in space and time". In: *Geological Society Special Publications* 121 (1997), pp. 19–37.

[1110] Giorgio Ranalli. "Rheology and deep tectonics". In: *Annals of Geophysics* 40.3 (1997).

[1111] M. Rancic, R.J. Purser, and F. Mesinger. "A global shallow-water model using an expanded spherical cube: Gnomonic versus conformal coordinates". In: *Q. J. R. Meteorol. Soc.* 122 (1996), pp. 959–982.

[1112] R. Rannacher and S. Turek. "Simple Nonconforming Quadrilateral Stokes Element". In: *Numerical Methods for Partial Differential Equations* 8 (1992), pp. 97–111. DOI: 10.1002/num.1690080202.

[1113] J.T. Ratcliff, G. Schubert, and A. Zebib. "Steady tetrahedral and cubic patterns of spherical shell convection with temperature-dependent viscosity". In: *J. Geophys. Res.: Solid Earth* 101.B11 (1996), pp. 25, 473–25, 484. DOI: 10.1029/96JB02097.

[1114] HT Rathod and Sridevi Kilari. "General complete Lagrange family for the cube in finite element interpolations". In: *Computer Methods in Applied Mechanics and Engineering* 181.1-3 (2000), pp. 295–344. DOI: 10.1016/S0045-7825(99)00080-8.

[1115] P-A Raviart and Jean-Marie Thomas. "Primal hybrid finite element methods for 2nd order elliptic equations". In: *Mathematics of computation* 31.138 (1977), pp. 391–413. DOI: 10.1090/S0025-5718-1977-0431752-8.

[1116] William H Raymond and Arthur Garder. "Selective damping in a Galerkin method for solving wave problems with variable grids". In: *Monthly weather review* 104.12 (1976), pp. 1583–1590.

[1117] J.N. Reddy. "On penalty function methods in the finite element analysis of flow problems". In: *Int. J. Num. Meth. Fluids* 2 (1982), pp. 151–171. DOI: 10.1002/fld.1650020204.

[1118] J.N. Reddy and D.K. Gartling. *The Finite Element Method in Heat Transfer and Fluid Dynamics.* CRC Press, 2010. ISBN: 978-1-4200-8598-3.

[1119] Hannah L Redmond and Scott D King. "A numerical study of a mantle plume beneath the Tharsis Rise: Reconciling dynamic uplift and lithospheric support models". In: *J. Geophys. Res.: Planets* 109.E9 (2004). DOI: 10.1029/2003JE002228.

[1120] K. Regenauer-Lieb and D.A. Yuen. "Rapid conversion of elastic energy into plastic shear heating during incipient necking of the lithosphere". In: *Geophys. Res. Lett.* 25.14 (1998), pp. 2737–2740. DOI: 10.1029/98GL02056.

[1121] Klaus Regenauer-Lieb and Jean-Pierre Petit. "Cutting of the European continental lithosphere: Plasticity theory applied to the present Alpine collision". In: *J. Geophys. Res.: Solid Earth* 102 (Apr. 1997), pp. 7731–7746.

[1122] Klaus Regenauer-Lieb et al. "From point defects to plate tectonic faults". In: *Philosophical Magazine* 86.21-22 (2006), pp. 3373–3392. DOI: 10.1080/14786430500375159.

[1123] Mirko Reguzzoni and Daniele Sampietro. "GEMMA: An Earth crustal model based on GOCE satellite data". In: *International Journal of Applied Earth Observation and Geoinformation* 35 (2015), pp. 31–43. DOI: 10.1016/j.jag.2014.04.002.

[1124] Mirko Reguzzoni and Daniele Sampietro. "Moho estimation using GOCE data: a numerical simulation". In: *Geodesy for Planet Earth.* Springer, 2012, pp. 205–214.

[1125] Mirko Reguzzoni, Daniele Sampietro, and Fernando Sansò. "Global Moho from the combination of the CRUST2. 0 model and GOCE data". In: *Geophy. J. Int.* 195.1 (2013), pp. 222–237.

[1126] M. ur Rehman, C. Vuik, and G. Segal. "SIMPLE-type preconditioners for the Oseen problem". In: *Int. J. Num. Meth. Fluids* 61 (2009), pp. 432–452. DOI: 10.1002/fld.1957.

[1127] Robert J Renka. "Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation package". In: *ACM Transactions on Mathematical Software (TOMS)* 22.1 (1996), pp. 1–8.

[1128] Robert J Renka. "Algorithm 772: STRIPACK: Delaunay triangulation and Voronoi diagram on the surface of a sphere". In: *ACM Transactions on Mathematical Software (TOMS)* 23.3 (1997), pp. 416–434.

[1129] J. Revenaugh and B. Parsons. "Dynamic topography and gravity anomalies for fluid layers whose viscosity varies exponentially with depth". In: *Geophysical Journal of the Royal Astronomical Society* 90.2 (1987), pp. 349–368. DOI: 10.1111/j.1365-246X.1987.tb00731.x.

[1130] Neil M Ribe. "Analytical approaches to mantle dynamics". In: *Vol 7: Mantle Dynamics - Treatise on Geodynamics* 7 (2007), pp. 167–226. DOI: 10.1016/B978-044452748-6.00117-6.

[1131] Neil M Ribe. "Mantle flow induced by back arc spreading". In: *Geophy. J. Int.* 98.1 (1989), pp. 85–91.

[1132] Neil M Ribe. *Theoretical Mantle Dynamics*. Cambridge University Press, 2018.

[1133] NM Ribe. "The dynamics of plume-ridge interaction - II. Off-ridge plumes". In: *J. Geophys. Res.: Solid Earth* 101.B7 (1996), pp. 16195–16204.

[1134] NM Ribe and UR Christensen. "Three-dimensional modeling of plume-lithosphere interaction". In: *J. Geophys. Res.: Solid Earth* 99.B1 (1994), pp. 669–682.

[1135] NM Ribe, UR Christensen, and J Theissing. "The dynamics of plume-ridge interaction, 1: Ridge-centered plumes". In: *Earth Planet. Sci. Lett.* 134.1-2 (1995), pp. 155–168.

[1136] Y. Ricard, L. Fleitout, and C. Froidevaux. "Geoid heights and lithospheric stresses for a dynamic Earth". In: *Annales Geophysicae* 2.3 (1984), pp. 267–286. DOI: xxxx.

[1137] FD Richards, MJ Hoggard, LR Cowton, and NJ White. "Reassessing the thermal structure of oceanic lithosphere with revised global inventories of basement depths and heat flow measurements". In: *J. Geophys. Res.: Solid Earth* 123.10 (2018), pp. 9136–9161. DOI: 10.1029/2018JB015998.

[1138] Mark A Richards and Bradford H Hager. "Geoid anomalies in a dynamic Earth". In: *J. Geophys. Res.: Solid Earth* 89.B7 (1984), pp. 5987–6002.

[1139] F. Richter and D. McKenzie. "Simple plate models of mantle convection". In: *J. Geophys.* 44 (1978), pp. 441–471. DOI: xxxx.

[1140] F.M. Richter and S.F. Daly. "Convection models having a multiplicy of large horizontal scales". In: *J. Geophys. Res.: Solid Earth* 83 (1978), pp. 4951–4956.

[1141] Frank M Richter. "Convection and the large-scale circulation of the mantle". In: *J. Geophys. Res.: Solid Earth* 78.35 (1973), pp. 8735–8745. DOI: 10.1029/JB078i035p08735.

[1142] Frank M Richter. "Dynamical models for sea floor spreading". In: *Reviews of Geophysics* 11.2 (1973), pp. 223–287. DOI: 10.1029/RG011i002p00223.

[1143] Frank M Richter and Dan P McKenzie. "Parameterizations for the horizontally averaged temperature of infinite Prandtl number convection". In: *J. Geophys. Res.: Solid Earth* 86.B3 (1981), pp. 1738–1744. DOI: 10.1029/JB086iB03p01738.

[1144] A. E. Ringwood. *Composition and Petrology of the Earth's Mantle*. McGraw-Hill, New York, 1975.

[1145] B. Riviere. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations*. SIAM, 2008.

[1146] GP Roberts, HA Barnes, and P Carew. "Modelling the flow behaviour of very shear-thinning liquids". In: *Chemical Engineering Science* 56.19 (2001), pp. 5617–5623.

[1147] J. H. Roberts and J. Arkani-Hamed. "Impact-induced mantle dynamics on Mars". In: *Icarus* 218.1 (2012), pp. 278–289. DOI: 10.1016/j.icarus.2011.11.038.

[1148] J. H. Roberts and S. Zhong. "Degree-1 convection in the Martian mantle and the origin of the hemispheric dichotomy". In: *J. Geophys. Res.: Solid Earth* 111.E6 (2006). DOI: 10.1029/2005JE002668.

[1149] J.M. Robey. "On the Design, Implementation, and Use of a Volume-of-Fluid Interface Tracking Algorithm For Modeling Convection and other Processes in the Earth's Mantle". PhD thesis. University of California Davis, 2019.

[1150] Jonathan M Robey and Elbridge Gerry Puckett. "Implementation of a volume-of-fluid method in a finite element code with applications to thermochemical convection in a density stratified fluid in the earth's mantle". In: *Computers & Fluids* 190 (2019), pp. 217–253.

[1151] M.P. Robichaud and P.A. Tanguy. "Comparison of 3-D finite elements for fluid flow". In: *Communications in applied numerical methods* 3 (1987), pp. 223–233.

[1152] M.P. Robichaud and P.A. Tanguy. "Finite element solution of three-dimensional incompressible fluid flow problems by a preconditioned conjugate residual method". In: *Int. J. Num. Meth. Eng.* 24 (1987), pp. 447–457.

[1153] Jörg Robl and Kurt Stüwe. "Continental collision with finite indenter strength: 1. Concept and model formulation". In: *Tectonics* 24.4 (2005).

[1154] T Rolf, Nicolas Coltice, and PJ Tackley. "Linking continental drift, plate tectonics and the thermal state of the Earth's mantle". In: *Earth Planet. Sci. Lett.* 351 (2012), pp. 134–146.

[1155] B. Romanowicz. "Can we resolve 3D density heterogeneity in the lower mantle". In: *Geophys. Res. Lett.* 28.6 (2001), pp. 1107–1110.

[1156] C. Ronchi, R. Iacono, and P.S. Paolucci. "The "Cubed Sphere": A New Method for the Solution of Partial Differential Equations in Spherical Geometry". In: *J. Comp. Phys.* 124 (1996), pp. 93–114.

[1157] I. Rose, B. Buffet, and T. Heister. "Stability and accuracy of free surface time integration in viscous flows". In: *Phys. Earth. Planet. Inter.* 262 (2017), pp. 90–100.

[1158] Vitoriano Ruas. "Mixed finite element methods with discontinuous pressures for the axisymmetric Stokes problem". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics* 83.4 (2003), pp. 249–264. DOI: 10.1002/zamm.200310032.

[1159] Johann Rudi, Yu-hsuan Shih, and Georg Stadler. "Advanced Newton methods for geodynamical models of Stokes flow with viscoplastic rheologies". In: *Geochem. Geophys. Geosyst.* 21.9 (2020), e2020GC009059. DOI: 10.1029/2020GC009059.

[1160] Johann Rudi, Georg Stadler, and Omar Ghattas. "Weighted BFBT preconditioner for Stokes flow problems with highly heterogeneous viscosity". In: *SIAM Journal on Scientific Computing* 39.5 (2017), S272–S297. DOI: 10.1137/16M108450X.

[1161] Johann Rudi et al. "An extreme-scale implicit solver for complex PDEs: highly heterogeneous flow in earth's mantle". In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*. ACM. 2015, p. 5. DOI: 10.1145/2807591.2807675.

[1162] Maxwell L Rudolph, Vedran Lekić, and Carolina Lithgow-Bertelloni. "Viscosity jump in Earth's mid-mantle". In: *Science* 350.6266 (2015), pp. 1349–1352. DOI: 10.1126/science.aad4972.

[1163] Thomas Ruedas, Paul J. Tackley, and Sean C. Solomon. "Thermal and compositional evolution of the martian mantle: Effects of phase transitions and melting". In: *Phys. Earth. Planet. Inter.* 216 (2013), pp. 32–58. DOI: 10.1016/j.pepi.2012.12.002.

[1164] Reiner Rummel, Weiyong Yi, and Claudia Stummer. "GOCE gravitational gradiometry". In: *Journal of Geodesy* 85.11 (2011), p. 777.

[1165] Y. Saad. "A flexible inner-outer preconditioned GMRES algorithm". In: *SIAM J. Sci. Comput.* 14.2 (1993), pp. 461–469.

[1166] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[1167] Y. Saad and M.H. Schultz. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems". In: *SIAM Journal on Scientific and Statistical Computing* 7.3 (1986), pp. 856–869. DOI: 10.1137/0907058.

[1168] M.H. Sadd. *Elasticity. Theory, applications and numerics (Third edition)*. Elsevier, 2014.

[1169] R. Sadourny. "Conservative Finite-Difference Approximations of the Primitive Equations on Quasi-Uniform Spherical Grids". In: *Monthly Weather Review* 100.2 (1972), pp. 136–144.

[1170] Edward B Saff and Amo BJ Kuijlaars. "Distributing many points on a sphere". In: *The mathematical intelligencer* 19.1 (1997), pp. 5–11. DOI: xxxx.

[1171] M. Saito and Y. Abe. "Consequences of anisotropic viscosity in the Earth's mantle (in Japanese, with English Abstract)". In: *Zisin* 37 (1984), pp. 237–245.

[1172] A. Salih. *Streamfunction-Vorticity Formulation*. Tech. rep. Department of Aerospace Engineering Indian Institute of Space Science and Technology, Thiruvananthapuram, 2013.

[1173] M. Sambridge, J. Braun, and H. McQueen. "Computational methods for natural neighbour interpolation in two and three dimensions". In: *Computational Techniques and Applications*. Ed. by R. L. May and A. K. Easton (eds). World Scientific, 1996, pp. 685–692.

[1174] M. Sambridge, J. Braun, and H. McQueen. "Geophysical parametrization and interpolation of irregular data using natural neighbours". In: *Geophy. J. Int.* 122 (1995), pp. 837–857. DOI: 10.1111/j.1365-246X.1995.tb06841.x.

[1175] Daniele Sampietro. "Geological units and Moho depth determination in the Western Balkans exploiting GOCE data". In: *Geophy. J. Int.* 202.2 (2015), pp. 1054–1063.

[1176] H Samuel, V Aleksandrov, and B Deo. "The effect of continents on mantle convective stirring". In: *Geophys. Res. Lett.* 38.4 (2011), p. L04307. DOI: 10.1029/2010GL046056.

[1177] H. Samuel and M. Evonuk. "Modeling advection in geophysical flows with particle level sets". In: *Geochem. Geophys. Geosyst.* 11.8 (2010). DOI: 10.1029/2010GC003081.

[1178] Henri Samuel. "A deformable particle-in-cell method for advective transport in geodynamic modelling". In: *Geophy. J. Int.* 214.3 (2018), pp. 1744–1773.

[1179] Henri Samuel, Maxim D Ballmer, Sebastiano Padovan, Nicola Tosi, Attilio Rivoldini, and Ana-catalina Plesa. "The Thermo-Chemical Evolution of Mars With a Strongly Stratified Mantle". In: *J. Geophys. Res.: Planets* 126.4 (2021), e2020JE006613. DOI: 10.1029/2020JE006613.

[1180] Henri Samuel and Nicola Tosi. "The influence of post-perovskite strength on the Earth's mantle thermal and chemical evolution". In: *Earth Planet. Sci. Lett.* 323 (2012), pp. 50–59.

[1181] Henri Samuel et al. "Geophysical evidence for an enriched molten silicate layer above Mars's core". In: *Nature* 622.7984 (2023), pp. 712–717. DOI: 10.1038/s41586-023-06601-8.

[1182] Patrick Sanan, Dave A May, Matthias Bollhöfer, and Olaf Schenk. "Pragmatic solvers for 3D Stokes and elasticity problems with heterogeneous coefficients: evaluating modern incomplete LDL T preconditioners". In: *Solid Earth* 11.6 (2020), pp. 2031–2045. DOI: 10.5194/se-11-2031-2020.

[1183] Patrick Sanan, Dave A May, Richard T Mills, et al. "DMStag: staggered, structured grids for PETSc". In: *Journal of Open Source Software* 7.79 (2022), p. 4531. DOI: 10.21105/joss.04531.

[1184] R.L. Sani, P.M. Gresho, R.L. Lee, and D.F. Griffiths. "The cause and cure (?) of the spurious pressures generated by certain FEM solutions of the incompressible Navier-Stokes equations: part 1". In: *Int. J. Num. Meth. Fluids* 1 (1981), pp. 17–43. DOI: 10.1002/fld.1650010104.

[1185] R.L. Sani, P.M. Gresho, R.L. Lee, D.F. Griffiths, and M. Engelman. "The cause and cure (?) of the spurious pressures generated by certain FEM solutions of the incompressible Navier-Stokes equations: part 2". In: *Int. J. Num. Meth. Fluids* 1 (1981), pp. 171–204. DOI: 10.1002/fld.1650010206.

[1186] Pierre Saramito. "A damped Newton algorithm for computing viscoplastic fluid flows". In: *Journal of Non-Newtonian fluid mechanics* 238 (2016), pp. 6–15. DOI: 10.1016/j.jnnfm.2016.05.007.

[1187] Pierre Saramito. "A new elastoviscoplastic model based on the Herschel–Bulkley viscoplastic model". In: *Journal of Non-Newtonian Fluid Mechanics* 158.1-3 (2009), pp. 154–161. DOI: 10.1016/j.jnnfm.2008.12.001.

[1188] Pierre Saramito. *Complex fluids*. Springer, 2016. ISBN: 978-3-319-44361-4.

[1189] Pierre Saramito and Anthony Wachs. "Progress in numerical simulation of yield stress fluid flows". In: *Rheologica Acta* 56.3 (2017), pp. 211–230. DOI: 10.1007/s00397-016-0985-9.

[1190] Atuo Sato and Erik G Thompson. "Finite element models for creeping convection". In: *J. Comp. Phys.* 22.2 (1976), pp. 229–244. DOI: 10.1016/0021-9991(76)90077-2.

[1191] Michael Schäfer, Stefan Turek, Franz Durst, Egon Krause, and Rolf Rannacher. "Benchmark computations of laminar flow around a cylinder". In: *Flow simulation with high-performance computers II*. 1996, pp. 547–566. DOI: 10.1007/978-3-322-89849-4_39.

[1192] S.M. Schmalholz. "A simple analytical solution for slab detachment". In: *Earth Planet. Sci. Lett.* 304 (2011), pp. 45–54. DOI: 10.1016/j.epsl.2011.01.011.

[1193] S.M. Schmalholz, Y.Yu. Podladchikov, and D.W. Schmid. "A spectral/finite difference method for simulating large deformations of heterogeneous, viscoelastic materials". In: *Geophy. J. Int.* 145 (2001), pp. 199–208. DOI: 10.1046/j.0956-540x.2000.01371.x.

[1194] Stefan M Schmalholz. "3D numerical modeling of forward folding and reverse unfolding of a viscous single-layer: Implications for the formation of folds and fold patterns". In: *Tectonophysics* 446.1-4 (2008), pp. 31–41.

[1195] J. Schmalzl and U. Hansen. "Mixing the Earth's mantle by thermal convection: A scale dependent phenomenon". In: *Geophys. Res. Lett.* 21.11 (1994), pp. 987–990. DOI: 10.1029/94GL00049.

[1196] J. Schmalzl, G.A. Houseman, and U. Hansen. "Mixing in vigorous, time-dependent three-dimensional convection and application to Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 101.B10 (1996), pp. 21847–21858.

[1197] J. Schmalzl, G.A. Houseman, and U. Hansen. "Mixing properties of three-dimensional (3-D) stationary convection". In: *Physics of Fluids* 7.5 (1995), pp. 1027–1033. DOI: 10.1063/1.868614.

[1198] Jörg Schmalzl and Alexander Loddoch. "Using subdivision surfaces and adaptive surface simplification algorithms for modeling chemical heterogeneities in geophysical flows". In: *Geochem. Geophys. Geosyst.* 4.9 (2003). DOI: 10.1029/2003GC000578.

[1199] H. Schmeling and W.R. Jacoby. "On modelling the lithosphere in mantle convection with non-linear rheology". In: *Journal of Geophysics* 50 (1981), pp. 89–100.

[1200] H. Schmeling et al. "A benchmark comparison of spontaneous subduction models - Towards a free surface". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 198–223. DOI: `10.1016/j.pepi.2008.06.028`.

[1201] Harro Schmeling. "Compressible convection with constant and variable viscosity: The effect on slab formation, geoid, and topography". In: *J. Geophys. Res.: Solid Earth* 94.B9 (1989), pp. 12463–12481.

[1202] Harro Schmeling. "On the relation between initial conditions and late stages of Rayleigh-Taylor instabilities". In: *Tectonophysics* 133.1-2 (1987), pp. 65–80. DOI: `10.1016/0040-1951(87)90281-2`.

[1203] Harro Schmeling, Alexander R Cruden, and Gabriele Marquart. "Finite deformation in and around a fluid sphere moving through a viscous medium: implications for diapiric ascent". In: *Tectonophysics* 149.1-2 (1988), pp. 17–34. DOI: `10.1016/0040-1951(88)90116-3`.

[1204] D.W. Schmid and Y.Y. Podlachikov. "Analytical solutions for deformable elliptical inclusions in general shear". In: *Geophy. J. Int.* 155 (2003), pp. 269–288. DOI: `10.1046/j.1365-246X.2003.02042.x`.

[1205] Max W Schmidt and Stefano Poli. "Experimentally based water budgets for dehydrating slabs and consequences for arc magma generation". In: *Earth Planet. Sci. Lett.* 163.1-4 (1998), pp. 361–379. DOI: `10.1016/S0012-821X(98)00142-3`.

[1206] G.E. Schneider, G.D. Raithby, and M.M. Yovanovich. "Finite-element solution procedures for solving the incompressible Navier-Stokes equations using equal order variable interpolation". In: *Numerical Heat Transfer* 1 (1978), pp. 433–451.

[1207] Robert Schneiders. "A grid-based algorithm for the generation of hexahedral element meshes". In: *Engineering with computers* 12.3-4 (1996), pp. 168–177.

[1208] Robert Schneiders. "Algorithms for quadrilateral and hexahedral mesh generation". In: *Proceedings of the VKI Lecture Series on Computational Fluid Dynamic, VKI-LS* 4 (2000).

[1209] Robert Schneiders. "Quadrilateral and Hexahedral Element Meshes". In: *chapter 21?* (1999).

[1210] Robert Schneiders. "Refining quadrilateral and hexahedral element meshes". In: *transition* 2 (1996), p. 1.

[1211] Robert Schneiders and Jürgen Debye. "Refining quadrilateral and brick element meshes". In: *Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations*. 1995, pp. 53–65.

[1212] Martin PJ Schöpfer, Conrad Childs, and Tom Manzocchi. "Three-dimensional failure envelopes and the brittle-ductile transition". In: *J. Geophys. Res.: Solid Earth* 118.4 (2013), pp. 1378–1392. DOI: `10.1002/jgrb.50081`.

[1213] B Schott and H Schmeling. "Delamination and detachment of a lithospheric root". In: *Tectonophysics* 296.3-4 (1998), pp. 225–247. DOI: `10.1016/S0040-1951(98)00154-1`.

[1214] P. Schroeder and G. Lube. "Stabilised dG-FEM for incompressible natural convection flows with boundary and moving interior layers on non-adapted meshes". In: *J. Comp. Phys.* 335 (2017), pp. 760–779.

[1215] G Schubert, DL Turcotte, and ER Oxburgh. "Stability of planetary interiors". In: *Geophy. J. Int.* 18.5 (1969), pp. 441–460. DOI: `10.1111/j.1365-246X.1969.tb03370.x`.

[1216] G. Schubert, D.L. Turcotte, and P. Olson. *Mantle Convection in the Earth and Planets*. Cambridge University Press, 2001. ISBN: 0-521-70000-0. DOI: `10.1017/CBO9780511612879`.

[1217]   Gerald Schubert, Charles Anderson, and Peggy Goldman. "Mantle plume interaction with an endothermic phase change". In: *J. Geophys. Res.: Solid Earth* 100.B5 (1995), pp. 8245–8256. DOI: 10.1029/95JB00032.

[1218]   Gerald Schubert and Charles A Anderson. "Finite element calculations of very high Rayleigh number thermal convection". In: *Geophy. J. Int.* 80.3 (1985), pp. 575–601. DOI: 10.1111/j.1365-246X.1985.tb05112.x.

[1219]   Gerald Schubert and David A Yuen. "Shear heating instability in the Earth's upper mantle". In: *Tectonophysics* 50.2-3 (1978), pp. 197–205. DOI: 10.1016/0040-1951(78)90135-X.

[1220]   Gerald Schubert, David A Yuen, and Donald L Turcotte. "Role of phase transitions in a dynamic mantle". In: *Geophy. J. Int.* 42.2 (1975), pp. 705–735. DOI: 10.1111/j.1365-246X.1975.tb05888.x.

[1221]   Melchior Schuh-Senlis, Cedric Thieulot, Paul Cupillard, and Guillaume Caumon. "Towards the application of Stokes flow equations to structural restoration simulations". In: *Solid Earth* 11 (2020), pp. 1909–1930. DOI: 10.5194/se-11-1909-2020.

[1222]   P.R. Schunk, M.A. Heroux, R.R. Rao, T.A. Baer, S.R. Subia, and A.C. Sun. *Iterative solvers and preconditioners for fully-coupled finite element formulations of incompressible fluid mechanics and related transport problems*. Tech. rep. SAND2001-3512J. Sandia National Laboratories, 2001.

[1223]   Larkin Ridgway Scott and Michael Vogelius. "Conforming finite element methods for incompressible and nearly incompressible continua". In: *Lectures in Applied Mathematics* 22.2 (1985). DOI: xxxx.

[1224]   A. Segal. *Finite element methods for the incompressible Navier-Stokes equations*. Delft University of Technology, 2012.

[1225]   A. Segal, M. ur Rehman, and C. Vuik. "Preconditioners for Incompressible Navier-Stokes Solvers". In: *Numer. Math. Theor. Meth. Appl.* 3.3 (2010), pp. 245–275. DOI: 10.4208/nmtma.2010.33.1.

[1226]   P. Sekhar and S.D. King. "3D spherical models of Martian mantle convection constrained by melting history". In: *Earth Planet. Sci. Lett.* 388 (2014), pp. 27–37. DOI: 10.1016/j.epsl.2013.11.047.

[1227]   C. Echevarria Serur. "Fast iterative methods for solving the incompressible Navier-Stokes equations". PhD thesis. TU Delft, 2013.

[1228]   Ruben Sevilla and Thibault Duretz. "A face-centered finite volume method for high-contrast Stokes interface problems". In: *Int. J. Num. Meth. Eng.* 124 (2023), pp. 3709–3732. DOI: 10.1002/nme.7294.

[1229]   MH Shahnas and WR Peltier. "The impacts of mantle phase transitions and the iron spin crossover in ferropericlase on convective mixing - is the evidence for compositional convection definitive? New results from a Yin-Yang overset grid-based control volume model". In: *J. Geophys. Res.: Solid Earth* 120.8 (2015), pp. 5884–5910. DOI: 10.1002/2015JB012064.

[1230]   Farzin Shakib, Thomas JR Hughes, and Zdeněk Johan. "A new finite element formulation for computational fluid dynamics: X. The compressible Euler and Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 89.1-3 (1991), pp. 141–219.

[1231]   A. Shamekhi and A. Aliabadi. "Non-Newtonian lid-driven cavity flow simulation by mesh free method". In: *ICCES* 11.3 (2009), pp. 67–72.

[1232]   W. Sharples, L.N. Moresi, M. Velic, M.A. Jadamec, and D.A. May. "Simulating faults and plate boundaries with a transversely isotropic plasticity model". In: *Phys. Earth. Planet. Inter.* 252 (2016), pp. 77–90. DOI: 10.1016/j.pepi.2015.11.007.

[1233] Dongwoo Sheen. "P1–Nonconforming Polyhedral Finite Elements in High Dimensions". In: *2018 MATRIX Annals*. Ed. by Jan de Gier, Cheryl E. Praeger, and Terence Tao. Cham: Springer International Publishing, 2020, pp. 121–133. ISBN: 978-3-030-38230-8. DOI: `10.1007/978-3-030-38230-8_9`.

[1234] J.R. Shewchuk. "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain". In: (1994).

[1235] J.R. Shewchuk. "Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator". In: *Applied Computational Geometry: Towards Geometric Engineering*. Ed. by Ming C. Lin and Dinesh Manocha. Vol. 1148. Lecture Notes in Computer Science. From the First ACM Workshop on Applied Computational Geometry. 1996, pp. 203–222. DOI: `10.1007/BFb0014497`.

[1236] Jonathan Richard Shewchuk. "Delaunay refinement algorithms for triangular mesh generation". In: *Computational geometry* 22.1-3 (2002), pp. 21–74.

[1237] Jonathan Richard Shewchuk. "Reprint of: Delaunay refinement algorithms for triangular mesh generation". In: *Computational Geometry* 47.7 (2014), pp. 741–778. DOI: `10.1016/j.comgeo.2014.02.005`.

[1238] Yanan Shi, Dongping Wei, Zhong-Hai Li, Ming-Qi Liu, and Mengxue Liu. "Subduction mode selection during slab and mantle transition zone interaction: Numerical modeling". In: *Pure Appl. Geophys.* (2017), pp. 5–24. DOI: `10.1007/s00024-017-1762-0`.

[1239] J. S. Shiau, A.V. Lyamin, and S. W. Sloan. "Bearing capacity of a sand layer on clay by finite element limit analysis". In: *Can. Geotech. J.* 40 (2003), pp. 900–915.

[1240] M. Shipeng and S. Zhongci. "Nonconforming rotated Q_1 element on non-tensor product anisotropic meshes". In: *Science in China Series A: Mathematics* 49.10 (2006), pp. 1363–1375. DOI: `10.1007/s11425-006-1363-3`.

[1241] Hang Si. "TetGen, a Delaunay-based quality tetrahedral mesh generator". In: *ACM Transactions on Mathematical Software (TOMS)* 41.2 (2015), pp. 1–36. DOI: `10.1145/2629697`.

[1242] Karin Sigloch and Mitchell G Mihalynuk. "Intra-oceanic subduction shaped the assembly of Cordilleran North America". In: *Nature* 496.7443 (2013), p. 50.

[1243] Amit Kaur Sihota. "Conjugate gradient methods using MPI for distributed systems". MA thesis. McGill University, 2004.

[1244] Jesse L Silverberg, Matthew Bierbaum, James P Sethna, and Itai Cohen. "Collective motion of humans in mosh and circle pits at heavy metal concerts". In: *Physical review letters* 110.22 (2013), p. 228701. DOI: `10.1103/PhysRevLett.110.228701`.

[1245] D.J. Silvester and N. Kechkar. "Stabilised bilinear-constant velocity-pressure finite elements for the conjugate gradient solution of the stokes problem". In: *Computer Methods in Applied Mechanics and Engineering* 79.1 (1990), pp. 71–86. DOI: `10.1016/0045-7825(90)90095-4`.

[1246] Nathan Sime, Jakob M Maljaars, Cian R Wilson, and Peter E van Keken. "An exactly mass conserving and pointwise divergence free velocity method: Application to compositional buoyancy driven flow problems in geodynamics". In: *Geochem. Geophys. Geosyst.* 22.4 (2021), e2020GC009349. DOI: `10.1029/2020GC009349`.

[1247] Nathan Sime and Cian R Wilson. "Automatic weak imposition of free slip boundary conditions via Nitsche's method: application to nonlinear problems in geodynamics". In: *arXiv preprint arXiv:2001.10639* (2020).

[1248] Nathan Sime, Cian R Wilson, and Peter E van Keken. "A pointwise conservative method for thermochemical convection under the compressible anelastic liquid approximation". In: *Geochem. Geophys. Geosyst.* 23.2 (2022), e2021GC009922. DOI: `10.1029/2021GC009922`.

[1249] N.A. Simmons, S.C. Myers, G. Johannesson, and E. Matzel. "LLNL-G3Dv3: Global P wave tomography model for improved regional and teleseismic travel time prediction". In: *J. Geophys. Res.: Solid Earth* 117.B10302 (2012). DOI: 10.1029/2012JB009525.

[1250] Guy Simpson. *Practical Finite Element Modeling in Earth Science Using Matlab*. Wiley-Blackwell, 2017. ISBN: 978-1-119-24862-0.

[1251] N.H. Sleep, S. Stein, R.J. Geller, and R.G. Gordon. "Comment on "The use of the minimum-dissipation principle in tectonophysics"". In: *Earth Planet. Sci. Lett.* 45 (1979), pp. 218–220.

[1252] S.W. Sloan. "A fortran program for profile and wavefront reduction". In: *Int. J. Num. Meth. Eng.* 28 (1989), pp. 2651–2679.

[1253] S.W. Sloan. "An algorithm for profile and wavefront reduction of sparse matrices". In: *Int. J. Num. Meth. Eng.* 23.2 (1986), pp. 239–251.

[1254] Suzanne E Smrekar et al. "Pre-mission InSights on the interior of Mars". In: *Space Science Reviews* 215.1 (2019), pp. 1–72. DOI: 10.1007/s11214-018-0563-9.

[1255] F. Soboutia, A. Ghodsb, and J. Arkani-Hamed. "On the advection of sharp material interfaces in geodynamic problems: entrainment of the D" layer". In: *Journal of Geodynamics* 31 (2001), pp. 459–479.

[1256] Larry P Solheim and WR Peltier. "Avalanche effects in phase transition modulated thermal convection: A model of Earth's mantle". In: *J. Geophys. Res.: Solid Earth* 99.B4 (1994), pp. 6997–7018. DOI: 10.1029/93JB02168.

[1257] Larry P Solheim and WR Peltier. "Phase boundary deflections at 660-km depth and episodically layered isochemical convection in the mantle". In: *J. Geophys. Res.: Solid Earth* 99.B8 (1994), pp. 15861–15875. DOI: 10.1029/94JB00730.

[1258] LP Solheim and WR Peltier. "Heat transfer and the onset of chaos in a spherical, axisymmetric, anelastic model of whole mantle convection". In: *Geophysical & Astrophysical Fluid Dynamics* 53.4 (1990), pp. 205–255.

[1259] VS Solomatov. "Scaling of temperature-and stress-dependent viscosity convection". In: *Physics of Fluids* 7.2 (1995), pp. 266–274.

[1260] C Sotin and S Labrosse. "Three-dimensional thermal convection in an iso-viscous, infinite Prandtl number fluid heated from within and from below: applications to the transfer of heat through planetary mantles". In: *Phys. Earth. Planet. Inter.* 112.3-4 (1999), pp. 171–190. DOI: 10.1016/S0031-9201(99)00004-7.

[1261] A. Soulaimani, M. Fortin, Y. Ouellet, G. Dhatt, and F. Bertrand. "Simple continuous pressure elements for two- and three-dimensional incompressible flows". In: *Computer Methods in Applied Mechanics and Engineering* 62 (1987), pp. 47–69. DOI: 10.1016/0045-7825(87)90089-2.

[1262] M. Souli and J.P. Zolesio. "Arbitrary Lagrangian-Eulerian and free surface methods in fluid mechanics". In: *Computer Methods in Applied Mechanics and Engineering* 191 (2001), pp. 451–466. DOI: 10.1016/S0045-7825(01)00313-9.

[1263] E.A. de Souza Neto, D. Peric, and D.R.J. Owen. *Computational methods for plasticity*. Wiley, 2008.

[1264] Frank J Spera, David A Yuen, and Stephen J Kirschvink. "Thermal boundary layer convection in silicic magma chambers: Effects of temperature-dependent rheology and implications for thermogravitational chemical fractionation". In: *J. Geophys. Res.: Solid Earth* 87.B10 (1982), pp. 8755–8767. DOI: 10.1029/JB087iB10p08755.

[1265]   Charles G Speziale. "On the advantages of the vorticity-velocity formulation of the equations of fluid dynamics". In: *J. Comp. Phys.* 73 (1987), pp. 476–480.

[1266]   Edward A Spiegel and G Veronis. "On the Boussinesq approximation for a compressible fluid." In: *The Astrophysical Journal* 131 (1960), p. 442.

[1267]   M. Spiegelman, D.A. May, and C. Wilson. "On the solvability of incompressible Stokes with viscoplastic rheologies in geodynamics". In: *Geochem. Geophys. Geosyst.* 17 (2016), pp. 2213–2238. DOI: 10.1002/2015GC006228.

[1268]   Marc Spiegelman and Richard F Katz. "A semi-Lagrangian Crank-Nicolson algorithm for the numerical solution of advection-diffusion problems". In: *Geochem. Geophys. Geosyst.* 7.4 (2006), Q04014. DOI: 10.1029/2005GC001073.

[1269]   Marc Spiegelman and Dan McKenzie. "Simple 2-D models for melt extraction at mid-ocean ridges and island arcs". In: *Earth Planet. Sci. Lett.* 83.1-4 (1987), pp. 137–152.

[1270]   Frank D Stacey. "A thermal model of the Earth". In: *Phys. Earth. Planet. Inter.* 15.4 (1977), pp. 341–348.

[1271]   Frank D Stacey and Paul M Davis. *Physics of the Earth.* 2008.

[1272]   G. Stadler, M. Gurnis, C. Burstedde, L.C. Wilcox, L. Alisic, and O. Ghattas. "The dynamics of plate tectonics and mantle flow: from local to global scales". In: *Science* 329 (2010), pp. 1033–1038. DOI: 10.1126/science.1191223.

[1273]   Simon C. Stähler et al. "Seismic detection of the martian core". In: *Science* 373.6553 (2021), pp. 443–448. DOI: 10.1126/science.abi7730.

[1274]   Andrew Staniforth and Jean Côté. "Semi-Lagrangian integration schemes for atmospheric models - A review". In: *Monthly weather review* 119.9 (1991), pp. 2206–2223.

[1275]   Matthew Staten and Scott A Canann. "Post refinement element shape improvement for quadrilateral meshes". In: *220 Trends in Unstructured Mesh Generation, ASME.* Citeseer. 1997.

[1276]   Ph. Steer, R. Cattin, J. Lavé, and V. Godard. "Surface Lagrangian Remeshing: A new tool for studying long term evolution of continental lithosphere from 2D numerical modelling". In: *Computers and Geosciences* 37.8 (2011), pp. 1067–1074. DOI: 10.1016/j.cageo.2010.05.023.

[1277]   C Stein and U Hansen. "Arrhenius rheology versus Frank-Kamenetskii rheology - Implications for mantle dynamics". In: *Geochem. Geophys. Geosyst.* 14.8 (2013), pp. 2757–2770. DOI: 10.1002/ggge.20158.

[1278]   C. Stein, J. Lowman, and U. Hansen. "A comparison of mantle convection models featuring plates". In: *Geochem. Geophys. Geosyst.* 15 (2014), pp. 2689–2698. DOI: 10.1002/2013GC005211.

[1279]   S. Stein. "A model for the relation between spreading rate and oblique spreading". In: *Earth Planet. Sci. Lett.* 39 (1978), pp. 313–318.

[1280]   Volker Steinbach, Ulrich Hansen, and Adolf Ebel. "Compressible convection in the earth's mantle: a comparison of different approaches". In: *Geophys. Res. Lett.* 16.7 (1989), pp. 633–636. DOI: 10.1029/GL016i007p00633.

[1281]   Volker Steinbach and David A Yuen. "The non-adiabatic nature of mantle convection as revealed by passive tracers". In: *Earth Planet. Sci. Lett.* 136.3-4 (1995), pp. 241–250. DOI: 10.1016/0012-821X(95)00166-A.

[1282] Volker Steinbach, David A Yuen, and Wuling Zhao. "Instabilities from phase transitions and the timescales of mantle thermal evolution". In: *Geophys. Res. Lett.* 20.12 (1993), pp. 1119–1122. DOI: 10.1029/93GL01243.

[1283] B Steinberger and R Holme. "Mantle flow models with core-mantle boundary constraints and chemical heterogeneities in the lowermost mantle". In: *J. Geophys. Res.: Solid Earth* 113.B5 (2008).

[1284] B. Steinberger and A.R. Calderwood. "Models of large-scale viscous flow in the Earth's mantle with constraints from mineral physics and surface observations". In: *Geophy. J. Int.* 167 (2006), pp. 1461–1481. DOI: 10.1111/j.1365-246X.2006.03131.x.

[1285] Bernhard Steinberger. "Topography caused by mantle density variations: observation-based estimates and models derived from tomography and lithosphere thickness". In: *Geophy. J. Int.* 205.1 (2016), pp. 604–621. DOI: 10.1093/gji/ggw040.

[1286] Bernhard Steinberger, Stephanie C Werner, and Trond H Torsvik. "Deep versus shallow origin of gravity anomalies, topography and volcanism on Earth, Venus and Mars". In: *Icarus* 207.2 (2010), pp. 564–577. DOI: 10.1016/j.icarus.2009.12.025.

[1287] K. Stemmer, H. Harder, and U. Hansen. "A new method to simulate convection with strongly temperature- and pressure-dependent viscosity in a spherical shell: Applications to the Earth's mantle". In: *Phys. Earth. Planet. Inter.* 157 (2006), pp. 223–249. DOI: 10.1016/j.pepi.2006.04.007.

[1288] R. Stenberg. "Analysis of Mixed Finite Element Methods for the Stokes Problem: A unified approach". In: *Mathematics of Computation* 42.165 (1984), pp. 9–23.

[1289] R. Stenberg. "Error analysis of some finite element methods for the Stokes problem". In: *Mathematics of Computation* 54.190 (1990), pp. 495–508. DOI: 10.1090/S0025-5718-1990-1010601-X.

[1290] Rolf Stenberg. "On some three-dimensional finite elements for incompressible media". In: *Computer Methods in Applied Mechanics and Engineering* 63.3 (1987), pp. 261–269. DOI: 10.1016/0045-7825(87)90072-7.

[1291] Rolf Stenberg. "Some new families of finite elements for the Stokes equations". In: *Numerische Mathematik* 56 (1989), pp. 827–838. DOI: 10.1007/BF01405291.

[1292] Ove Stephansson and Harald Berner. "The finite element method in tectonic processes". In: *Phys. Earth. Planet. Inter.* 4.4 (1971), pp. 301–321.

[1293] Pietro Sternai. "Surface processes forcing on extensional rock melting". In: *Scientific Reports* 10.1 (2020), pp. 1–13. DOI: 10.1038/s41598-020-63920-w.

[1294] Eli Sternberg and F Rosenthal. "The elastic sphere under concentrated loads". In: *J. Appl. Mech.* 19(4) (1952), pp. 413–421. DOI: 10.1115/1.4010536.

[1295] D. Sterpi. "An analysis of geotechnical problems involving strain softening effects". In: *International Journal fro Numerical and Analytical Methods in Geomechanics* 23 (1999), pp. 1427–1454. DOI: 10.1002/(SICI)1096-9853(199911)23:13<1427::AID-NAG6>3.0.CO;2-B.

[1296] WNR Stevens. "Finite element, stream function–vorticity solution of steady laminar natural convection". In: *Int. J. Num. Meth. Fluids* 2.4 (1982), pp. 349–366. DOI: 10.1002/fld.1650020404.

[1297] Lars Stixrude and Carolina Lithgow-Bertelloni. "Mineralogy and elasticity of the oceanic upper mantle: Origin of the low-velocity zone". In: *J. Geophys. Res.: Solid Earth* 110.B3 (2005).

[1298] Lars Stixrude and Carolina Lithgow-Bertelloni. "Thermodynamics of mantle minerals – I. Physical properties". In: *Geophy. J. Int.* 162.2 (2005), pp. 610–632. DOI: 10.1111/j.1365-246X.2005.02642.x.

[1299] Glen S Stockmal, Christopher Beaumont, and Ross Boutilier. "Geodynamic models of convergent margin tectonics: transition from rifted margin to overthrust belt and consequences for foreland-basin development". In: *AAPG Bulletin* 70.2 (1986), pp. 181–190. DOI: 10.1306/94885656-1704-11D7-8645000102C1865D.

[1300] Yvonne Marie Stokes. "Very Viscous Flows Driven by Gravity with particular application to Slumping of Molten Glass". PhD thesis. University of Adelaide, 1998.

[1301] J. Suckale, B.H. Hager, L.T. Elkins-Tanton, and J.-Ch. Nave. "It takes three to tango: 2. Bubble dynamics in basaltic volcanoes and ramifications for modeling normal Strombolian activity". In: *J. Geophys. Res.: Solid Earth* 115.B7 (2010). DOI: 10.1029/2009JB006917.

[1302] J. Suckale, J.-C. Nave, and B.H. Hager. "It takes three to tango: 1. Simulating buoyancy-driven flow in the presence of large viscosity contrasts". In: *J. Geophys. Res.: Solid Earth* 115 (2010), B07409. DOI: 10.1029/2009JB006916.

[1303] J. Sung, H.G. Choi, and J.Y. Yoo. "Time-accurate computation of unsteady free surface flows using an ALE-segregated equal-order FEM". In: *Computer Methods in Applied Mechanics and Engineering* 190 (2000), pp. 1425–1440.

[1304] M. Sussman and E.G. Puckett. "A Coupled Level Set and Volume-of-Fluid Method for Computing 3D and Axisymmetric Incompressible Two-Phase Flows". In: *J. Comp. Phys.* 162 (2000), pp. 301–337.

[1305] K. Sverdrup, N. Nikiforakis, and A. Almgren. "Highly parallelisable simulations of time-dependent viscoplastic fluid flow simulations with structured adaptive mesh refinement". In: *Physics of Fluids* 30 (2018), p. 093102. DOI: 10.1063/1.5049202.

[1306] R. Swinbank and R.J. Purser. "Fibonacci grids: A novel approach to global modelling". In: *Quarterly Journal of the Royal Meteorological Society* 132 (2006), pp. 1769–1793.

[1307] A. Syrakos, G.C. Georgiou, and A.N. Alexandrou. "Performance of the finite volume method in solving regularised Bingham flows: Inertia effects in the lid-driven cavity flow". In: *Journal of Non-Newtonian Fluid Mechanics* 208–209 (2014), pp. 88–107. DOI: 10.1016/j.jnnfm.2014.03.004.

[1308] M. Tabata and A. Suzuki. "A stabilized finite element method for the Rayleigh-Bénard equations with infinite Prandtl number in a spherical shell". In: *Computer Methods in Applied Mechanics and Engineering* 190 (2000), pp. 387–402.

[1309] Masahisa Tabata. "Finite element analysis of axisymmetric flow problems". In: *Zeitschrift für angewandte Mathematik und Mechanik* 76 (1996), pp. 171–174.

[1310] M. Tachibana and Y. Iemoto. "Steady flow around, and drag on a circular cylinder moving at low speeds in a viscous liquid between two parallel planes". In: *Fluid Dynamics Research* 2 (1987), pp. 125–137.

[1311] P. Tackley. "Three-dimensional models of mantle convection: Influence of phase transitions and temperature-dependent viscosity". PhD thesis. California Institute of Technology, 1994, p290.

[1312] P.J. Tackley. "Modelling compressible mantle convection with large viscosity contrasts in a three-dimensional spherical shell using the yin-yang grid". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 7–18.

[1313] P.J. Tackley and S.D. King. "Testing the tracer ratio method for modeling active compositional fields in mantle convection simulations". In: *Geochem. Geophys. Geosyst.* 4.4 (2003). DOI: 10.1029/2001GC000214.

[1314] Paul J Tackley and Shunxing Xie. "The thermochemical structure and evolution of Earth's mantle: constraints and numerical models". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360.1800 (2002), pp. 2593–2609. DOI: 10.1098/rsta.2002.1082.

[1315] Xue–Cheng Tai and Ragnar Winther. "A discrete de Rham complex with enhanced smoothness". In: *Calcolo* 43 (2006), pp. 287–306. DOI: 10.1007/s10092-006-0124-6.

[1316] E. Taliadorou, G.C. Georgiou, and I. Moulitsas. "Weakly compressible Poiseuille flows of a Herschel-Bulkley fluid". In: *Journal of Non-Newtonian Fluid Mechanics* 158 (2009), pp. 162–169. DOI: 10.1016/j.jnnfm.2008.11.010.

[1317] M. Talwani and M. Ewing. "Rapid Computation of Gravitational Attraction of Three-Dimensional Bodies of Arbitrary Shape". In: *Geophysics* 25.1 (1960), pp. 203–225.

[1318] Manik Talwani, J Lamar Worzel, and Mark Landisman. "Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone". In: *J. Geophys. Res.: Solid Earth* 64.1 (1959), pp. 49–59.

[1319] E. Tan and M. Gurnis. "Compressible thermochemical convection and application to lower mantle structures". In: *J. Geophys. Res.: Solid Earth* 112.B06304 (2007).

[1320] A. Tanaka, Y. Sanada, Y. Yamada, T. Matsuoka, and Y. Ashida. "Discrete element simulations of continental collision in Asia". In: *Exploration Geophysics* 36.1 (2005), pp. 1–6. DOI: 10.1071/EG05001.

[1321] R.I. Tanner and E. Tanner. "Heinrich Hencky: a rheological pioneer". In: *Rheol. Acta* 42 (2003), pp. 93–101.

[1322] P. Tapponnier, G. Peltzer, A.Y. Le Dain, R. Armijo, and P. Cobbold. "Propagating extrusion tectonics in Asia: new insights from simple experiments with plasticine". In: *Geology* 10 (1982), pp. 611–616. DOI: 10.1130/0091-7613(1982)10<611:PETIAN>2.0.CO;2.

[1323] Paul Tapponnier and Peter Molnar. "Slip-line field theory and large-scale continental tectonics". In: *Nature* 264 (Nov. 1976), pp. 319–324. DOI: 10.1038/264319a0.

[1324] J. Tarduno, H.P. Bunge, N. Sleep, and U. Hansen. "The bent hawaiian-emperor hotspot track: inheriting the mantle wind". In: *Science* 324.5923 (2009), pp. 50–53. DOI: 10.1126/science.1161256.

[1325] Alexandra J. Taylor and Simon D. R. Wilson. "Conduit flow of an incompressible, yield-stress fluid". In: *Journal of Rheology* 41 (1997), pp. 93–101. DOI: 10.1122/1.550802.

[1326] C. Taylor and P. Hood. "A numerical solution of the Navier-Stokes equations using the finite element technique". In: *Comput. Fluids* 1 (1973), pp. 73–100. DOI: 10.1016/0045-7930(73)90027-3.

[1327] Arkady Ten, David A Yuen, Tine B Larsen, and Andrei V Malevsky. "The evolution of material surfaces in convection with variable viscosity as monitored by a characteristics-based method". In: *Geophys. Res. Lett.* 23.16 (1996), pp. 2001–2004. DOI: 10.1029/96GL02182.

[1328] Arkady Ten, David A Yuen, Yu Yu Podladchikov, Tine B Larsen, Elizaveta Pachepsky, and Andrei V Malevsky. "Fractal features in mixing of non-Newtonian and Newtonian mantle convection". In: *Earth Planet. Sci. Lett.* 146.3-4 (1997), pp. 401–414. DOI: 10.1016/S0012-821X(96)00244-0.

[1329] Arkady A Ten, Yuri Yu Podladchikov, David A Yuen, Tine B Larsen, and Andrei V Malevsky. "Comparison of mixing properties in convection with the Particle-Line Method". In: *Geophys. Res. Lett.* 25.16 (1998), pp. 3205–3208. DOI: https://doi.org/10.1029/98GL51991.

[1330] Andy R Terrel, L Ridgway Scott, Matthew Gregg Knepley, Robert C Kirby, and Garth N Wells. "Finite elements for incompressible fluids". In: *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book.* Ed. by Anders Logg, Kent-Andre Mardal, and Garth Wells. Vol. 84. Springer, 2012, pp. 385–397. DOI: 10.1007/978-3-642-23099-8_20.

[1331] J.L. Tetreault and S.J.H. Buiter. "Geodynamic models of terrane accretion: Testing the fate of island arcs, oceanic plateaus, and continental fragments in subduction zones". In: *J. Geophys. Res.: Solid Earth* 117 (2012), B08403. DOI: 10.1029/2012JB009316.

[1332] Jannis Teunissen and Ute Ebert. "Afivo: A framework for quadtree/octree AMR with shared-memory parallelization and geometric multigrid methods". In: *Computer Physics Communications* 233 (2018), pp. 156–166.

[1333] T.E. Tezduyar. "Stabilized Finite Element Formulations for Incompressible Flow Computations". In: *Advances in Applied Mechanics* 28 (1992), pp. 1–44.

[1334] Tayfun E Tezduyar, Mittal Behr, S Mittal, and J Liou. "A new strategy for finite element computations involving moving boundaries and interfaces - the deforming-spatial-domain/space-time procedure: II. Computation of free-surface flows, two-liquid flows, and flows with drifting cylinders". In: *Computer Methods in Applied Mechanics and Engineering* 94.3 (1992), pp. 353–371.

[1335] Tayfun E Tezduyar, Sanjay Mittal, SE Ray, and R Shih. "Incompressible flow computations with stabilized bilinear and linear equal-order-interpolation velocity-pressure elements". In: *Computer Methods in Applied Mechanics and Engineering* 95.2 (1992), pp. 221–242.

[1336] Tayfun E. Tezduyar and Yasuo Osawa. "Finite element stabilization parameters computed from element matrices and vectors". In: *Computer Methods in Applied Mechanics and Engineering* 190 (2000), pp. 411–430. DOI: 10.1016/S0045-7825(00)00211-5.

[1337] TE Tezduyar and YJ Park. "Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations". In: *Computer Methods in Applied Mechanics and Engineering* 59.3 (1986), pp. 307–325.

[1338] Thomas M Tharp. "Numerical models of subduction and forearc deformation". In: *Geophy. J. Int.* 80.2 (1985), pp. 419–437. DOI: 10.1111/j.1365-246X.1985.tb05102.x.

[1339] M. Thielmann, B.J.P. Kaus, and A.A. Popov. "Lithospheric stresses in Rayleigh-Benard convection: effects of a free surface and a viscoelastic Maxwell rheology". In: *Geophy. J. Int.* 203 (2015), pp. 2200–2219. DOI: 10.1093/gji/ggv436.

[1340] M. Thielmann, D.A. May, and B.J.P. Kaus. "Discretization errors in the Hybrid Finite Element Particle-In-Cell Method". In: *Pure Appl. Geophys.* 171.9 (2014), pp. 2164–2184. DOI: 10.1007/s00024-014-0808-9.

[1341] C. Thieulot. "Analytical solution for viscous incompressible Stokes flow in a spherical shell". In: *Solid Earth* 8 (2017), pp. 1181–1191. DOI: 10.5194/se-8-1181-2017.

[1342] C. Thieulot. "ELEFANT: a user-friendly multipurpose geodynamics code". In: *Solid Earth Discussions* 6 (2014), pp. 1949–2096. DOI: 10.5194/sed-6-1949-2014.

[1343] C. Thieulot. "FANTOM: two- and three-dimensional numerical modelling of creeping flows for the solution of geological problems". In: *Phys. Earth. Planet. Inter.* 188.1 (2011), pp. 47–68. DOI: 10.1016/j.pepi.2011.06.011.

[1344] C. Thieulot. "GHOST: Geoscientific Hollow Sphere Tesselation". In: *Solid Earth* 9.1–9 (2018). DOI: 10.5194/se-9-1169-2018.

[1345] C. Thieulot and W. Bangerth. "On the choice of finite element for applications in geodynamics". In: *Solid Earth* 13 (2022), pp. 229–249. DOI: 10.5194/se-13-229-2022.

[1346] C. Thieulot and W. Bangerth. "On the choice of finite element for applications in geodynamics. Part II: A comparison of simplex and hypercube elements". In: *Solid Earth* 16 (2025), pp. 457–476. DOI: 10.5194/se-16-457-2025.

[1347] C. Thieulot, P. Fullsack, and J. Braun. "Adaptive octree-based finite element analysis of two- and three-dimensional indentation problems". In: *J. Geophys. Res.: Solid Earth* 113 (2008), B12207. DOI: 10.1029/2008JB005591.

[1348] C. Thieulot, P. Steer, and R.S. Huismans. "Three-dimensional numerical simulations of crustal systems undergoing orogeny and subjected to surface processes". In: *Geochem. Geophys. Geosyst.* 15 (2014). DOI: 10.1002/2014GC005490.

[1349] Cedric Thieulot and Pep Español. "Non-isothermal diffusion in a binary mixture with smoothed particle hydrodynamics". In: *Computer physics communications* 169.1-3 (2005), pp. 172–176. DOI: 10.1016/j.cpc.2005.03.039.

[1350] Cedric Thieulot, LPBM Janssen, and Pep Español. "Smoothed particle hydrodynamics model for phase separating fluid mixtures. I. General equations". In: *Physical Review E* 72.1 (2005), p. 016713. DOI: 10.1103/PhysRevE.72.016713.

[1351] Cedric Thieulot, LPBM Janssen, and Pep Español. "Smoothed particle hydrodynamics model for phase separating fluid mixtures. II. Diffusion in a binary mixture". In: *Physical Review E* 72.1 (2005), p. 016714. DOI: 10.1103/PhysRevE.72.016714.

[1352] B Thomas, H Samuel, CG Farnetani, J Aubert, and C Chauvel. "Mixing time of heterogeneities in a buoyancy-dominated magma ocean". In: *Geophy. J. Int.* 236.2 (2024), pp. 764–777. DOI: 10.1093/gji/ggad452.

[1353] Paul J Thomas and Steven W Squyres. "Formation of crater palimpsests on Ganymede". In: *J. Geophys. Res.: Solid Earth* 95.B12 (1990), pp. 19161–19174. DOI: 10.1029/JB095iB12p19161.

[1354] Paul J Thomas and Steven W Squyres. "Relaxation of impact basins on icy satellites". In: *J. Geophys. Res.: Solid Earth* 93.B12 (1988), pp. 14919–14932. DOI: 10.1029/JB093iB12p14919.

[1355] Paul J Thomas, Steven W Squyres, and Michael H Carr. "Flank tectonics of Martian volcanoes". In: *J. Geophys. Res.: Solid Earth* 95.B9 (1990), pp. 14345–14355. DOI: 10.1029/JB095iB09p14345.

[1356] J.F. Thompson, B.K. Soni, and N.P. Weatherill. *Handbook of grid generation.* CRC press, 1998.

[1357] F. Tin-Loi and N.S. Ngo. "Performance of the p-version finite element method for limit analysis". In: *International Journal of Mechanical Sciences* 45 (2003), pp. 1149–1166. DOI: 10.1016/j.ijmecsci.2003.08.004.

[1358] C. Tirel, J.-P. Brun, and E. Burov. "Dynamics and structural development of metamorphic core complexes". In: *J. Geophys. Res.: Solid Earth* 113.B04403 (2008).

[1359] Max Tirone. "On the thermal gradient in the Earth's deep interior". In: *Solid Earth* 7.1 (2016), p. 229. DOI: 10.5194/se-7-229-2016.

[1360] Andréa Tommasi, Mickael Knoll, Alain Vauchez, Javier W Signorelli, Catherine Thoraval, and Roland Logé. "Structural reactivation in plate tectonics controlled by olivine crystal anisotropy". In: *Nature Geoscience* 2.6 (2009), pp. 423–427. DOI: 10.1038/ngeo528.

[1361] Rosaria Tondi, Maddalena Gilardoni, and Mirko Reguzzoni. "The combined inversion of seismological and GOCE gravity data: New insights into the current state of the Pacific lithosphere and upper mantle". In: *Tectonophysics* 705 (2017), pp. 101–115. DOI: `10.1016/j.tecto.2017.03.013`.

[1362] KE Torrance and DL Turcotte. "Structure of convection cells in the mantle". In: *J. Geophys. Res.: Solid Earth* 76.5 (1971), pp. 1154–1161. DOI: `10.1029/JB076i005p01154`.

[1363] KE Torrance and DL Turcotte. "Thermal convection with large viscosity variations". In: *Journal of Fluid Mechanics* 47.1 (1971), pp. 113–125. DOI: `10.1017/S002211207100096X`.

[1364] A. Toselli and O. Widlund. *Domain decomposition methods - Algorithms and Theory.* Springer, 2005.

[1365] N. Tosi et al. "A community benchmark for viscoplastic thermal convection in a 2-D square box". In: *Geochem. Geophys. Geosyst.* 16.7 (2015), pp. 2175–2196. DOI: `10.1002/2015GC005807`.

[1366] Nicola Tosi. "Numerical modeling of present-day mantle convection". PhD thesis. Charles University in Prague, 2007.

[1367] Aaron Tovish, Gerald Schubert, and Bruce P Luyendyk. "Mantle flow pressure and the angle of subduction: Non-Newtonian corner flows". In: *J. Geophys. Res.: Solid Earth* 83.B12 (1978), pp. 5892–5898.

[1368] B.J. Travis et al. "A benchmark comparison of numerical methods for infinite Prandtl number thermal convection in two-dimensional Cartesian geometry". In: *Geophysical & Astrophysical Fluid Dynamics* 55.3-4 (1990), pp. 137–160.

[1369] Bryan Travis and Peter Olson. "Convection with internal heat sources and thermal turbulence in the Earth's mantle". In: *Geophy. J. Int.* 118.1 (1994), pp. 1–19. DOI: `10.1111/j.1365-246X.1994.tb04671.x`.

[1370] Bryan Travis, Peter Olson, and Gerald Schubert. "The transition from two-dimensional to three-dimensional planforms in infinite-Prandtl-number thermal convection". In: *Journal of Fluid Mechanics* 216 (1990), pp. 71–91. DOI: `10.1017/S0022112090000349`.

[1371] S.J. Trim, J. P. Lowman, and S.L. Butler. "Improving Mass Conservation With the Tracer Ratio Method: Application to Thermochemical Mantle Flows". In: *Geochem. Geophys. Geosyst.* 22 (2020), e2019GC008799. DOI: `10.1029/2019GC008799`.

[1372] Sean J Trim, Samuel L Butler, and Raymond J Spiteri. "Benchmarking multiphysics software for mantle convection". In: *Computers and Geosciences* 154 (2021), p. 104797. DOI: `10.1016/j.cageo.2021.104797`.

[1373] R.A. Trompert and U. Hansen. "On the Rayleigh number dependence of convection with a strongly temperature-dependent viscosity". In: *Physics of Fluids* 10.2 (1998), pp. 351–360.

[1374] R.A. Trompert and U. Hansen. "The application of a finite volume multigrid method to three-dimensional flow problems in a highly viscous fluid with a variable viscosity". In: *Geophysical & Astrophysical Fluid Dynamics* 83.3-4 (1996), pp. 261–291. DOI: `10.1080/03091929608208968`.

[1375] Ulrich Trottenberg, Cornelius W. Oosterlee, and Anton Schuller. *Multigrid.* Elsevier, 2001.

[1376] VP Trubitsyn, IE Rogozhina, and MK Kaban. "On a spectral method of solving the Stokes equation". In: *Izvestiya, Physics of the Solid Earth* 44.1 (2008), pp. 18–25.

[1377] Terry E Tullis, Franklin G Horowitz, and Jan Tullis. "Flow laws of polyphase aggregates from end-member flow laws". In: *J. Geophys. Res.: Solid Earth* 96.B5 (1991), pp. 8081–8096. DOI: `10.1029/90JB02491`.

[1378]   D.L. Turcotte and G. Schubert. *Geodynamics, 2nd edition.* Cambridge, 2012. ISBN: 9780511807442. DOI: 10.1017/CBO9780511807442.

[1379]   D.L. Turcotte and G. Schubert. *Geodynamics, 3nd edition.* Cambridge University Press, 2014. ISBN: 9780521186230,

[1380]   DL Turcotte and ER Oxburgh. "Finite amplitude convective cells and continental drift". In: *Journal of Fluid Mechanics* 28.1 (1967), pp. 29–42. DOI: 10.1017/S0022112067001880.

[1381]   S. Turek. "A comparative study of time-stepping techniques for the incompressible Navier-Stokes equations: from fully implicit non-linear schemes to semi-implicit projection methods". In: *Int. J. Num. Meth. Fluids* 22 (1996), pp. 987–1011.

[1382]   S. Turek. *Efficient Solvers for Incompressible Flow Problems.* Springer, Berlin, Heidelberg, 1999. ISBN: 978-3-540-65433-9. DOI: 10.1007/978-3-642-58393-3.

[1383]   S. Turek. "Tools for simulating non-stationary incompressible flow via discretely divergence-free finite element models". In: *Int. J. Num. Meth. Fluids* 18 (1994), pp. 71–105.

[1384]   S. Turek and A. Ouazzi. "Unified edge-oriented stabilization of nonconforming finite element methods for incompressible flow problems". In: *Journal of Numerical Mathematics* 15.4 (2007).

[1385]   Stefan Turek, Abderrahim Ouazzi, and Rainer Schmachtel. "Multigrid methods for stabilized nonconforming finite elements for incompressible flow involving the deformation tensor formulation". In: *Journal of Numerical Mathematics* 10.3 (2002), pp. 235–248.

[1386]   L. Uieda, V.C.F. Barbosa, and C. Braitenberg. "Tesseroids: Forward-modeling gravitational fields in spherical coordinates". In: *Geophysics* 81.5 (2015), pp. 41–48. DOI: 10.1190/GEO2015-0204.1.

[1387]   Leonardo Uieda. "Forward modeling and inversion of gravitational fields in spherical coordinates". PhD thesis. Observatorio Nacional, 2016.

[1388]   Leonardo Uieda, Everton P Bomfim, Carla Braitenberg, and Eder Molina. "Optimal forward calculation method of the Marussi tensor due to a geologic structure at GOCE height". In: *Proceedings of the 4th International GOCE User Workshop.* Munich Germany. 2011.

[1389]   H.C. Upadhyaya, O.P. Sharma, R. Mittal, and H. Fatima. "Icosahedral-hexagonal grids on a sphere for CFD applications". In: *Asia-Pac. J. Chem. Eng.* 6 (2011), pp. 110–119. DOI: 10.1002/apj.479.

[1390]   G de Vahl Davis and IP Jones. "Natural convection in a square cavity: a comparison exercise". In: *International Journal for numerical methods in fluids* 3.3 (1983), pp. 227–248. DOI: 10.1002/fld.1650030304.

[1391]   JL Valera, Ana M Negredo, and Ivone Jiménez-Munt. "Deep and near-surface consequences of root removal by asymmetric continental delamination". In: *Tectonophysics* 502.1-2 (2011), pp. 257–265. DOI: 10.1016/j.tecto.2010.04.002.

[1392]   Juan-Luis Valera, Ana-María Negredo, and Antonio Villaseñor. "Asymmetric delamination and convective removal numerical modeling: comparison with evolutionary models for the Alboran Sea region". In: *Pure Appl. Geophys.* 165 (2008), pp. 1683–1706. DOI: 10.1007/s00024-008-0395-8.

[1393]   FN Van de Vosse et al. "Finite-element-based computational methods for cardiovascular fluid-structure interaction". In: *Journal of engineering mathematics* 47.3-4 (2003), pp. 335–368. DOI: 10.1023/B:ENGI.0000007985.17625.43.

[1394]   A.P. Van Den Berg and D.A. Yuen. "The role of shear heating in lubricating mantle flow". In: *Earth Planet. Sci. Lett.* 151.1-2 (1997), pp. 33–42. DOI: 10.1016/S0012-821X(97)00110-6.

[1395]   A.P. van den Berg and D.A. Yuen. "Modelling planetary dynamics by using the temperature at the core-mantle boundary as a control variable: effects of rheological layering on mantle heat transport". In: *Phys. Earth. Planet. Inter.* 108.3 (1998), pp. 219–234. DOI: 10.1016/S0031-9201(98)00101-0.

[1396]   Arie P van den Berg, Peter E van Keken, and David A Yuen. "The effects of a composite non-Newtonian and Newtonian rheology on mantle convection". In: *Geophy. J. Int.* 115.1 (1993), pp. 62–78. DOI: 10.1111/j.1365-246X.1993.tb05588.x.

[1397]   Arie P van den Berg and David A Yuen. "Is the lower-mantle rheology Newtonian today?" In: *Geophys. Res. Lett.* 23.16 (1996), pp. 2033–2036. DOI: 10.1029/96GL02065.

[1398]   SP Van Der Pijl, A Segal, and C Vuik. "Modelling of three-dimensional bubbly flows with a mass-conserving Level-Set method". In: *Proceedings of the 4th European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS.* 2004.

[1399]   SP Van der Pijl, A Segal, C Vuik, and P Wesseling. "A mass-conserving level-set method for modelling of multi-phase flows". In: *International journal for numerical methods in fluids* 47.4 (2005), pp. 339–361. DOI: 10.1002/d.817.

[1400]   SP Van der Pijl, A Segal, C Vuik, and P Wesseling. "Computing three-dimensional two-phase flows with a mass-conserving level set method". In: *Computing and Visualization in Science* 11.4-6 (2008), pp. 221–235. DOI: 10.1007/s00791-008-0106-0.

[1401]   E. van der Wiel, D.J.J. van Hinsbergen, C. Thieulot, and W. Spakman. "Linking rates of slab sinking to long-term lower mantle flow and mixing". In: *Earth Planet. Sci. Lett.* 625 (2024), p. 118471. DOI: 10.1016/j.epsl.2023.118471.

[1402]   HJ van Heck and PJ Tackley. "Plate tectonics on super-Earths: equally or more likely than on Earth". In: *Earth Planet. Sci. Lett.* 310.3-4 (2011), pp. 252–261. DOI: 10.1016/j.epsl.2011.07.029.

[1403]   J. van Hunen, A.P. van den Berg, and N.J. Vlaar. "On the role of subducting oceanic plateaus in the development of shallow flat subduction". In: *Tectonophysics* 352.3-4 (2002), pp. 317–333. DOI: 10.1016/S0040-1951(02)00263-9.

[1404]   P. van Keken, D.A. Yuen, and A. van den Berg. "Pulsating diapiric flows: Consequences of vertical variations in mantle creep laws". In: *Earth Planet. Sci. Lett.* 112 (1992), pp. 179–194. DOI: 10.1016/0012-821X(92)90015-N.

[1405]   P. van Keken and S. Zhong. "Mixing in a 3D spherical model of present-day mantle convection". In: *Earth Planet. Sci. Lett.* 171 (1999), pp. 533–547. DOI: 10.1016/S0012-821X(99)00181-8.

[1406]   P.E. van Keken, S.D. King, H. Schmeling, U.R. Christensen, D. Neumeister, and M.-P. Doin. "A comparison of methods for the modeling of thermochemical convection". In: *J. Geophys. Res.: Solid Earth* 102.B10 (1997), pp. 22, 477–22, 495.

[1407]   P.E. van Keken, C.J. Spiers, A.P. van den Berg, and E.J. Muyzert. "The effective viscosity of rocksalt: implementation of steady-state creep laws in numerical models of salt diapirism". In: *Tectonophysics* 225 (1993), pp. 457–476. DOI: 10.1016/0040-1951(93)90310-G.

[1408]   P.E. van Keken, D.A. Yuen, and A.P. van den Berg. "The effects of shallow rheological boundaries in the upper mantle on inducing shorter time scales of diapiric flows". In: *Geophys. Res. Lett.* 20.18 (1993), pp. 1927–1930. DOI: 10.1029/93GL01768.

[1409]   P.E. van Keken et al. "A community benchmark for subduction zone modelling". In: *Phys. Earth. Planet. Inter.* 171 (2008), pp. 187–197. DOI: 10.1016/j.pepi.2008.04.015.

[1410]   P. van Thienen. "Convective vigour and heat flow in chemically differentiated systems". In: *Geophy. J. Int.* 169.2 (2007), pp. 747–766. DOI: 10.1111/j.1365-246X.2007.03377.x.

[1411]   R.S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Inc., 1963.

[1412]   Oleg V. Vasilyev, Yuri Yu. Podladchikov, and David A. Yuen. "Modelling of viscoelastic plume-lithosphere interaction using the adaptive multilevel wavelet collocation method". In: *Geophy. J. Int.* 147 (2001), pp. 579–589.

[1413]   J. Vatteville, P.E. van Keken, A. Limare, and A. Davaille. "Starting laminar plumes: Comparison of laboratory and numerical modeling". In: *Geochem. Geophys. Geosyst.* 10.12 (2009). DOI: 10.1029/2009GC002739.

[1414]   A. Vauchez, A. Tomassi, and G. Barroul. "Rheological heterogeneity, mechanical anisotropy and deformation of the continental lithosphere". In: *Tectonophysics* 296 (1998), pp. 61–86. DOI: 10.1016/S0040-1951(98)00137-1.

[1415]   M Velić, L. Moresi, D. May, and M. Knepley. "A Family of Numerically Stable Analytic Solutions for Geodynamic Code Verification". In: ().

[1416]   Mirko Velić, Dave May, and Louis Moresi. "A fast robust algorithm for computing discrete voronoi diagrams". In: *Journal of Mathematical Modelling and Algorithms* 8.3 (2009), pp. 343–355. DOI: 10.1007/s10852-008-9097-6.

[1417]   Fabio Verbosio, Arne De Coninck, Drosos Kourounis, and Olaf Schenk. "Enhancing the scalability of selected inversion factorization algorithms in genomic prediction". In: *Journal of Computational Science* 22.Supplement C (2017), pp. 99–108. ISSN: 1877-7503.

[1418]   Jean Verhoogen. "The adiabatic gradient in the mantle". In: *Eos, Transactions American Geophysical Union* 32.1 (1951), pp. 41–43.

[1419]   A Verruijt. "Deformations of an elastic half plane with a circular cavity". In: *International Journal of Solids and Structures* 35.21 (1998), pp. 2795–2804. DOI: 10.1016/S0020-7683(97)00194-7.

[1420]   Valérie Vidal. "Interaction des Différentes échelles de Convection dans le Manteau Terrestre". PhD thesis. Institut de physique du globe de paris-IPGP, 2004.

[1421]   Jean-Pierre Vilotte, M Daignieres, and Raul Madariaga. "Numerical modeling of intraplate deformation: Simple mechanical models of continental collision". In: *J. Geophys. Res.: Solid Earth* 87.B13 (1982), pp. 10709–10728. DOI: 10.1029/JB087iB13p10709.

[1422]   Jean-Pierre Vilotte, Marc Daignieres, R Madariaga, and OC Zienkiewicz. "The role of a heterogeneous inclusion during continental collision". In: *Phys. Earth. Planet. Inter.* 36.3-4 (1984), pp. 236–259. DOI: 10.1016/0031-9201(84)90049-9.

[1423]   Jean-Pierre Vilotte, Raul Madariaga, Marc Daignieres, and O Zienkiewicz. "Numerical study of continental collision: influence of buoyancy forces and an initial stiff inclusion". In: *Geophy. J. Int.* 84.2 (1986), pp. 279–310. DOI: 10.1111/j.1365-246X.1986.tb04357.x.

[1424]   A.P. Vincent and D.A. Yuen. "Thermal attractor in chaotic convection with high-Prandtl-number fluids". In: *Physical Review A* 38.1 (1988), pp. 328–334. DOI: 10.1103/PhysRevA.38.328.

[1425]   C. Vincent and R. Boyer. "A preconditioned conjugate gradient Uzawa-type method for the solution of the Stokes problem by mixed Q1-P0 stabilised finite elements". In: *Int. J. Num. Meth. Fluids* 14 (1992), pp. 289–298.

[1426]   N.J. Vlaar, P.E. van Keken, and A.P. van den Berg. "Cooling of the Earth in the Archean: Consequences of pressure-release melting in a hotter mantle". In: *Earth Planet. Sci. Lett.* 121 (1994), pp. 1–18. DOI: 10.1016/0012-821X(94)90028-0.

[1427] Michael Vogelius. "A right-inverse for the divergence operator in spaces of piecewise polynomials: Application to the p-version of the finite element method". In: *Numerische Mathematik* 41.1 (1983), pp. 19–37. DOI: 10.1007/BF01396303.

[1428] M. von Tscharner and S. M. Schmalholz. "A 3-D Lagrangian finite element algorithm with remeshing for simulating large-strain hydrodynamic instabilities in power law viscoelastic fluids". In: *Geochem. Geophys. Geosyst.* 16.1 (2015), pp. 215–245. DOI: 10.1002/2014GC005628.

[1429] H.A. van der Vorst and C. Vuik. "GMRESR: a family of nested GMRES methods". In: *Num. Lin. Alg. Appl.* 1 (1994), pp. 369–386.

[1430] C Vuik, A Saghir, and GP Boerstoel. "The Krylov accelerated SIMPLE (R) method for flow problems in industrial furnaces". In: *Int. J. Num. Meth. Fluids* 33.7 (2000), pp. 1027–1040.

[1431] Kees Vuik. "Een kwart eeuw iteratieve methoden". In: (2009).

[1432] L. Vynnytska, M.E. Rognes, and S.R. Clark. "Benchmarking FEniCS for mantle convection simulations". In: *Computers and Geosciences* 50 (2013), pp. 95–105. DOI: 10.1016/j.cageo.2012.05.012.

[1433] Uwe Walzer and Roland Hendel. "A new convection–fractionation model for the evolution of the principal geochemical reservoirs of the Earth's mantle". In: *Phys. Earth. Planet. Inter.* 112.3-4 (1999), pp. 211–256. DOI: 10.1016/S0031-9201(99)00035-7.

[1434] Uwe Walzer and Roland Hendel. "Tectonic episodicity and convective feedback mechanisms". In: *Phys. Earth. Planet. Inter.* 100.1-4 (1997), pp. 167–188. DOI: 10.1016/S0031-9201(96)03238-4.

[1435] Bo Wang and BC Khoo. "Hybridizable discontinuous Galerkin method (HDG) for Stokes interface flow". In: *J. Comp. Phys.* 247 (2013), pp. 262–278. DOI: 10.1016/j.jcp.2013.03.064.

[1436] H. Wang, R. Agrusta, and J. van Hunen. "Advantages of a conservative velocity interpolation (CVI) scheme for particle-in-cell methods with application in geodynamic modeling". In: *Geochem. Geophys. Geosyst.* 16 (2015). DOI: 10.1002/2015GC005824.

[1437] Li Wang and Dimitri J Mavriplis. "Adjoint-based h–p adaptive discontinuous Galerkin methods for the 2D compressible Euler equations". In: *J. Comp. Phys.* 228.20 (2009), pp. 7643–7661.

[1438] W.M. Wang, L.J. Sluys, and R. de Borst. "Viscoplasticity for instabilities due to strain softening and strain-rate softening". In: *Int. J. Num. Meth. Eng.* 40 (1997), pp. 3839–3864.

[1439] X. Wang and W.K. Liu. "Extended immersed boundary method using FEM and RKPM". In: *Computer Methods in Applied Mechanics and Engineering* 193 (2004), pp. 1305–1321. DOI: 10.1016/j.cma.2003.12.024.

[1440] Xinguo Wang, William E Holt, and Attreyee Ghosh. "Joint modeling of lithosphere and mantle dynamics: Evaluation of constraints from global tomography models". In: *J. Geophys. Res.: Solid Earth* 120.12 (2015), pp. 8633–8655. DOI: 10.1002/2015JB012188.

[1441] G.H. Wannier. "A contribution to the hydrodynamics of lubrication". In: *Quarterly of Applied Mathematics* VIII (1950), pp. 1–32.

[1442] C.J. Warren, C. Beaumont, and R.A. Jamieson. "Formation and exhumation of ultra-high-pressure rocks during continental collision: Role of detachment in the subduction channel". In: *Geochem. Geophys. Geosyst.* 9 (2008). DOI: 10.1029/2007GC001839.

[1443] C.J. Warren, C. Beaumont, and R.A. Jamieson. "Modelling tectonic styles and ultra-high pressure (UHP) rock exhumation during the transition from oceanic subduction to continental collision". In: *Earth Planet. Sci. Lett.* 267 (2008), pp. 129–145.

[1444] Clare J Warren, Christopher Beaumont, and Rebecca A Jamieson. "Deep subduction and rapid exhumation: Role of crustal strength and strain weakening in continental subduction and ultrahigh-pressure rock exhumation". In: *Tectonics* 27.6 (2008).

[1445] R.F. Weinberg and H. Schmeling. "Polydiapirs: multiwavelength gravity structures". In: *Journal of Structural Geology* 14.4 (1992), pp. 425–436. DOI: 10.1016/0191-8141(92)90103-4.

[1446] Theodore Weinberg and B Sousedík. "Fast implementation of mixed RT0 finite elements in MATLAB". In: *SIAM Undergraduate Research Online* 12 (2019). DOI: 10.1137/18S017430.

[1447] S.A. Weinstein, P.L. Olson, and D.A. Yuen. "Time-dependent large aspect-ratio thermal convection in the earth's mantle". In: *Geophysical & Astrophysical Fluid Dynamics* 47.1-4 (1989), pp. 157–197. DOI: 10.1080/03091928908221820.

[1448] M. B. Weller and A. Lenardic. "The Energetics and Convective Vigor of Mixed-mode Heating: Velocity Scalings and Implications for the Tectonics of Exoplanets: The Energetics of Mixed-mode convection". In: *Geophys. Res. Lett.* 43 (2016), pp. 9469–9474. DOI: 10.1002/2016GL069927.

[1449] M.B. Weller, A. Lenardic, and W.B. Moore. "Scaling relationships and physics for mixed heating convection in planetary interiors: Isoviscous spherical shells". In: *J. Geophys. Res.: Solid Earth* 121 (2016), pp. 7598–7617. DOI: 10.1002/2016JB013247.

[1450] R.A. Werner and D.J. Scheeres. "Exterior gravitation of a polyhedron derived and compared with harmonic and mascon gravitation representations of asteroid 4769 Castalia". In: *Celestial Mechanics and Dynamical Astronomy* 65 (1997), pp. 313–344. DOI: 10.1007/BF00053511.

[1451] Pieter Wesseling and Cornelis W Oosterlee. "Geometric multigrid with applications to computational fluid dynamics". In: *Journal of Computational and Applied Mathematics* 128.1-2 (2001), pp. 311–334. DOI: 10.1016/S0377-0427(00)00517-3.

[1452] D.M. Whipp, C. Beaumont, and J. Braun. "Feeding the "aneurysm": Orogen-parallel mass transport into Nanga Parbat and the western Himalayan syntaxis". In: *J. Geophys. Res.: Solid Earth* 119 (2014). DOI: 10.1002/2013JB010929.

[1453] A Whittaker, MHP Bott, and GD Waghorn. "Stresses and plate boundary forces associated with subduction plate margins". In: *J. Geophys. Res.: Solid Earth* 97.B8 (1992), pp. 11933–11944.

[1454] Mark A Wieczorek and Matthias Meschede. "Shtools: Tools for working with spherical harmonics". In: *Geochem. Geophys. Geosyst.* 19.8 (2018), pp. 2574–2592. DOI: 10.1029/2018GC007529.

[1455] Erik van der Wiel, Cedric Thieulot, and Douwe JJ van Hinsbergen. "Quantifying mantle mixing through configurational Entropy". In: *Solid Earth* 15 (2024), pp. 861–875. DOI: 10.5194/se-15-861-2024.

[1456] William SD Wilcock and JA Whitehead. "The Rayleigh-Taylor instability of an embedded layer of low-viscosity fluid". In: *J. Geophys. Res.: Solid Earth* 96.B7 (1991), pp. 12193–12200. DOI: 10.1029/91JB00339.

[1457] M.L. Wilkins. *Computer simulation of dynamic phenomena*. Springer, 1999.

[1458] K.R. Wilks and N.L. Carter. "Rheology of some continental lower crustal rocks". In: *Tectonophysics* 182.1-2 (1990), pp. 57–77. DOI: 10.1016/0040-1951(90)90342-6.

[1459] S.D. Willett. "Dynamic and kinematic growth and change of a Coulomb wedge". In: *Thrust Tectonics*. Ed. by K.R. McClay. Chapman and Hall, 1992, pp. 19–31.

[1460] S.D. Willett. "Orogeny and orography: The effects of erosion on the structure of mountain belts". In: *J. Geophys. Res.: Solid Earth* 104.B12 (1999), p. 28957.

[1461] S. Williams, L. Oliker, R. Vuduc, J. Shalf, K. Yelick, and J. Demmel. "Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms". In: *SC'07 proceedings of the 2007 ACM/IEEE conference on supercomputing*. ACM NY, 2007.

[1462] Samuel Williams et al. "PERI-auto-tuning memory-intensive kernels for multicore". In: *Journal of Physics: Conference Series*. Vol. 125. 1. IOP Publishing. 2008, p. 012038. DOI: 10.1088/1742-6596/125/1/012038.

[1463] Erskine D Williamson and Leason H Adams. "Density distribution in the Earth". In: *Journal of the Washington Academy of Sciences* 13.19 (1923), pp. 413–428. DOI: xxxx.

[1464] J Tuzo Wilson. "Did the Atlantic close and then re-open?" In: *Nature* 211.5050 (1966), pp. 676–681. DOI: 10.1038/211676a0.

[1465] HH Winter. "Viscous dissipation term in energy equations". In: *Calculation and Measurement Techniques for Momentum, Energy and Mass Transfer* 7 (1987), pp. 27–34. DOI: xxxx.

[1466] W.-D. Woidt. "Finite element calculations applied to salt dome analysis". In: *Tectonophysics* 50 (1978), pp. 369–386. DOI: 10.1016/0040-1951(78)90143-9.

[1467] Marek Wojciechowski. "A note on the differences between Drucker-Prager and Mohr-Coulomb shear strength criteria". In: *Studia Geotechnica et Mechanica* (2018). DOI: 10.2478/sgem-2018-0016.

[1468] M. Wolstencroft, J.H. Davies, and D.R. Davies. "Nusselt-Rayleigh number scaling for spherical shell Earth mantle simulation up to a Rayleigh number of 10^9". In: *Phys. Earth. Planet. Inter.* 176 (2009), pp. 132–141. DOI: 10.1016/j.pepi.2009.05.002.

[1469] Jennifer Worthen, Georg Stadler, Noemi Petra, Michael Gurnis, and Omar Ghattas. "Towards adjoint-based inversion for rheological parameters in nonlinear viscous mantle flow". In: *Phys. Earth. Planet. Inter.* 234 (2014), pp. 23–34.

[1470] Daniel G Wright. "Finite difference approximations to the advection-diffusion equation". In: *Tellus A* 44.3 (1992), pp. 261–269. DOI: 10.1034/j.1600-0870.1992.t01-2-00005.x.

[1471] G.B. Wright, N. Flyer, and D.A. Yuen. "A hybrid radial basis function-pseudospectral method for thermal convection in a 3-D spherical shell". In: *Geochem. Geophys. Geosyst.* 11.7 (2010). DOI: 10.1029/2009GC002985.

[1472] B. Wu, Z. Xu, and Z. Li. "A note on imposing displacement boundary conditions in finite element analysis". In: *Commun. Numer. Meth. Engng* 24 (2008), pp. 777–784.

[1473] Patrick Wu. "Deformation of an incompressible viscoelastic flat earth with powerlaw creep: a finite element approach". In: *Geophy. J. Int.* 108.1 (1992), pp. 35–51. DOI: 10.1111/j.1365-246X.1992.tb00837.x.

[1474] Patrick Wu and WR Peltier. "Viscous gravitational relaxation". In: *Geophy. J. Int.* 70.2 (1982), pp. 435–485. DOI: 10.1111/j.1365-246X.1982.tb04976.x.

[1475] H. Xing, W. Yu, and J. Zhang. "3D Mesh Generation in Geocomputing". In: *Advances in Geocomputing, Lecture Notes in Earth Sciences*. Berlin Heidelberg: Springer-Verlag, 2009. DOI: 10.1007/978-3-540-85879-9_2.

[1476] Dongbin Xiu and George Em Karniadakis. "A semi-Lagrangian high-order method for Navier–Stokes equations". In: *J. Comp. Phys.* 172.2 (2001), pp. 658–684. DOI: 10.1006/jcph.2001.6847.

[1477] Y. Lina nd Y. Cao. "A new nonlinear Uzawa algorithm for generalised saddle point problems". In: *Applied Mathematics and Computation* 175 (2006), pp. 1432–1454.

[1478] P. Yamato, L. Husson, J. Braun, C. Loiselet, and C. Thieulot. "Influence of surrounding plates on 3D subduction dynamics". In: *Geophys. Res. Lett.* 36.L07303 (2009). DOI: 10.1029/2008GL036942.

[1479] Haibin Yang, Louis Moresi, and John Mansour. "Stress Recovery for the Particle-in-cell Finite Element Method". In: *Phys. Earth. Planet. Inter.* 311 (2021), p. 106637. DOI: 10.1016/j.pepi.2020.106637.

[1480] S. Yang and Y. Shi. "Three-dimensional numerical simulation of glacial trough forming process". In: *Science China: Earth Sciences* (2015). DOI: 10.1007/s11430-015-5120-8.

[1481] Ting Yang, Louis Moresi, Dapeng Zhao, Dan Sandiford, and Joanne Whittaker. "Cenozoic lithospheric deformation in Northeast Asia and the rapidly-aging Pacific Plate". In: *Earth Planet. Sci. Lett.* 492 (2018), pp. 1–11. DOI: 10.1016/j.epsl.2018.03.057.

[1482] Woo-Sun Yang and John R Baumgardner. "A matrix-dependent transfer multigrid method for strongly variable viscosity infinite Prandtl number thermal convection". In: *Geophysical & Astrophysical Fluid Dynamics* 92.3-4 (2000), pp. 151–195. DOI: 10.1080/03091920008203715.

[1483] K. Yasuda, R.C. Armstrong, and R.E. Cohen. "Shear flow properties of concentrated solutions of linear and star branched polystyrenes". In: *Rheol. Acta* 20 (1981), pp. 163–178. DOI: 10.1007/BF01513059.

[1484] Irad Yavneh. "Why multigrid methods are so efficient". In: *Computing in science & engineering* 8.6 (2006), p. 12. DOI: 10.1109/MCSE.2006.125.

[1485] Tao Ye, Rajat Mittal, HS Udaykumar, and Wei Shyy. "An accurate Cartesian grid method for viscous incompressible flows with complex immersed boundaries". In: *J. Comp. Phys.* 156.2 (1999), pp. 209–240. DOI: 10.1006/jcph.1999.6356.

[1486] Liang Yin, Chao Yang, Shi-Zhuang Ma, and Ke-Ke Zhang. "Parallel and fully implicit simulations of the thermal convection in the Earth's outer core". In: *Computers & Fluids* 193 (2019), p. 104278. DOI: 10.1016/j.compfluid.2019.104278.

[1487] He Yiqian and Yang Haitian. "Solving inverse couple-stress problems via an element-free Galerkin (EFG) method and Gauss–Newton algorithm". In: *Finite Elements in Analysis and Design* 46.3 (2010), pp. 257–264. DOI: 10.1016/j.finel.2009.09.009.

[1488] M. Yoshida and A. Kageyama. "Application of the Yin-Yang grid to a thermal convection of a Boussinesq fluid with infinite Prandtl number in a three-dimensional spherical shell". In: *Geophys. Res. Lett.* 31.L12609 (2004). DOI: 10.1029/2004GL019970.

[1489] Masaki Yoshida, Satoru Honda, Motoyuki Kido, and Yasuyuki Iwase. "Numerical simulation for the prediction of the plate motions". In: *Earth, planets and space* 53.7 (2001), pp. 709–721. DOI: 10.1186/BF03352399.

[1490] Masaki Yoshida and Akira Kageyama. "Low-degree mantle convection with strongly temperature- and depth-dependent viscosity in a three-dimensional spherical shell". In: *J. Geophys. Res.: Solid Earth* 111.B3 (2006). DOI: 10.1029/2005JB003905.

[1491] S. Yoshioka and M.J.R. Wortel. "Three-dimensional numerical modeling of detachment of subducted lithosphere". In: *J. Geophys. Res.: Solid Earth* 100.B10 (1995), pp. 20, 223–20, 244.

[1492] DL Youngs. *Numerical methods for fluid dynamics, ed. by KW Morton and MJ Baines.* Academic Press, Massachusetts, 1982.

[1493] Hongzheng Yu and Shimin Wang. "Unified Linear Stability Analysis for Thermal Convections in Spherical Shells Under Different Boundary Conditions and Heating Modes". In: *Earth and Space Science* 6.9 (2019), pp. 1749–1768. DOI: 10.1029/2019EA000672.

[1494] X. Yu and F. Tin-Loi. "A simple mixed finite element for static limit analysis". In: *Computers and Structures* 84 (2006), pp. 1906–1917.

[1495] K.Y. Yuan, Y.S. Huang, H.T. Yang, and T.H.H. Pian. "The inverse mapping and distortion measures for 8-node hexahedral isoparametric elements". In: *Computational Mechanics* 14 (1994), pp. 189–199. DOI: 10.1007/BF00350284.

[1496] D.A. Yuen and W.R. Peltier. "Mantle plumes and the thermal stability of the D" layer". In: *Geophys. Res. Lett.* 7.9 (1980), pp. 625–628. DOI: 10.1029/GL007i009p00625.

[1497] D.A. Yuen, W.R. Peltier, and G. Schubert. "On the existence of a second scale of convection in the upper mantle". In: *Geophysical Journal of the Royal Astronomical Society* 65.1 (1981), pp. 171–190. DOI: 10.1111/j.1365-246X.1981.tb02707.x.

[1498] David A Yuen, Ctirad Matyska, Ondrej Cadek, and Masanori Kameyama. "The dynamical influences from physical properties in the lower mantle and post-perovskite phase transition". In: *Geophysical Monograph - American Geophysical Union* 174 (2007), p. 249. DOI: xxxx.

[1499] David A Yuen, DM Reuteler, S Balachandar, V Steinbach, AV Malevsky, and JJ Smedsmo. "Various influences on three-dimensional mantle convection with phase transitions". In: *Phys. Earth. Planet. Inter.* 86.1-3 (1994), pp. 185–203. DOI: 10.1016/0031-9201(94)05068-6.

[1500] S.T. Zalesak. "Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids". In: *J. Comp. Phys.* 31 (1979), pp. 335–362. DOI: 10.1016/0021-9991(79)90051-2.

[1501] S. Zaleski and P. Julien. "Numerical simulation of Rayleigh-Taylor instability for single and multiple salt diapirs". In: *Tectonophysics* 206 (1992), pp. 55–69. DOI: 10.1016/0040-1951(92)90367-F.

[1502] A. Zebib. "Linear and Weakly Nonlinear Variable Viscosity Convection in Spherical Shells". In: *Theoret. Comput. Fluid Dynamics* 4 (1993), pp. 241–253.

[1503] Abdelfattah Zebib, Gerald Schubert, and Joe M Straus. "Infinite Prandtl number thermal convection in a spherical shell". In: *Journal of Fluid Mechanics* 97.2 (1980), pp. 257–277. DOI: 10.1017/S0022112080002558.

[1504] Iris van Zelst, Fabio Crameri, Adina E. Pusok, Anne Glerum, Julianne Dannberg, and Cedric Thieulot. "101 Geodynamic modelling: How to design, interpret, and communicate numerical studies of the solid Earth". In: *Solid Earth* 13 (2022), pp. 583–637. DOI: 10.5194/se-13-583-2022.

[1505] Guoxiang Zhang and Junyu Xiang. "Eight-node conforming straight-side quadrilateral element with high-order completeness (QH8-C1)". In: *Int. J. Num. Meth. Eng.* (2020). DOI: 10.1002/nme.6360.

[1506] Huai Zhang, Lili Ju, Max Gunzburger, Todd Ringler, and Stephen Price. "Coupled models and parallel simulations for three-dimensional full-Stokes ice sheet modeling". In: *Numerical Mathematics: Theory, Methods and Applications* 4.3 (2011), pp. 396–418. DOI: 10.1017/S1004897900000416.

[1507] N. Zhang, S. Zhong, and A.K. McNarmara. "Supercontinent formation from stochastic collision and mantle convection models". In: *Gondwana Research* 15 (2009), pp. 267–275. DOI: 10.1016/j.gr.2008.10.002.

[1508] Shangyou Zhang. "A Family of $Q_{k+1,k} \times Q_{k,k+1}$ Divergence-Free Finite Elements on Rectangular Grids". In: *SIAM journal on numerical analysis* 47.3 (2009), pp. 2090–2107. DOI: 10.1137/080728949.

[1509] Shangyou Zhang. "A new family of stable mixed finite elements for the 3D Stokes equations". In: *Mathematics of Computation* 74 (2005), pp. 543–554. DOI: 10.1090/S0025-5718-04-01711-9.

[1510] Shangyou Zhang. "Divergence-free finite elements on tetrahedral grids for $k \geq 6$". In: *Mathematics of computation* 80.274 (2011), pp. 669–695. DOI: `10.1090/S0025-5718-2010-02412-3`.

[1511] Shangyou Zhang. "On the convergence of 2D P_4^+ triangular and 3D P_6^+ tetrahedral divergence-free finite elements". In: *Numerical Methods for Partial Differential Equations* 40.4 (2024), e23088. DOI: `10.1002/num.23088`.

[1512] Shangyou Zhang. "On the P1 Powell-Sabin divergence-free finite element for the Stokes equations". In: *Journal of Computational Mathematics* (2008), pp. 456–470. DOI: `xxx`.

[1513] Shangyou Zhang. "Quadratic divergence-free finite elements on Powell–Sabin tetrahedral grids". In: *Calcolo* 48 (2011), pp. 211–244. DOI: `10.1007/s10092-010-0035-4`.

[1514] Shangyou Zhang and Mo Mu. "Stable $Q_k - Q_{k-1}$ mixed finite elements with discontinuous pressure". In: *Journal of Computational and Applied Mathematics* 301 (2016), pp. 188–200. DOI: `10.1016/j.cam.2016.01.030`.

[1515] Yiping Zhang, Sizhao Wang, and Dave Chan. "A new five-node locking-free quadrilateral element based on smoothed FEM for near-incompressible linear elasticity". In: *Int. J. Num. Meth. Eng.* 100.9 (2014), pp. 633–668.

[1516] L. Zhao, X. Bai, T. Li, and J.J.R. Williams. "Improved conservative level set method". In: *Int. J. Num. Meth. Fluids* (2014). DOI: `10.1002/fld.3907`.

[1517] S. Zhong and M. Gurnis. "Role of plates and temperature-dependent viscosity in phase change dynamics". In: *J. Geophys. Res.: Solid Earth* 99.B8 (1994), p. 15903. DOI: `10.1029/94JB00545`.

[1518] S. Zhong, M. Gurnis, and G. Hulbert. "Accurate determination of surface normal stress in viscous flow from a consistent boundary flux method". In: *Phys. Earth. Planet. Inter.* 78.1-2 (1993), pp. 1–8. DOI: `10.1016/0031-9201(93)90078-N`.

[1519] S. Zhong, M. Gurnis, and L. Moresi. "Free-surface formulation of mantle convection-I. Basic theory and application to plumes". In: *Geophy. J. Int.* 127.3 (1996), pp. 708–718. DOI: `10.1111/j.1365-246X.1996.tb04049.x`.

[1520] S. Zhong, M. Gurnis, and L. Moresi. "Role of faults, nonlinear rheology, and viscosity structure in generating plates from instantaneous mantle flow models". In: *J. Geophys. Res.: Solid Earth* 103.B7 (1998), pp. 15, 255–15, 268. DOI: `10.1029/98JB00605`.

[1521] S. Zhong, A. McNamara, E. Tan, L. Moresi, and M. Gurnis. "A benchmark study on mantle convection in a 3-D spherical shell using CitcomS". In: *Geochem. Geophys. Geosyst.* 9.10 (2008). DOI: `10.1029/2008GC002048`.

[1522] S. Zhong and A.B. Watts. "Lithospheric deformation induced by loading of the Hawaiian Islands and its implications for mantle rheology". In: *J. Geophys. Res.: Solid Earth* 118 (2013), pp. 6025–6048. DOI: `10.1002/2013JB010408`.

[1523] S. Zhong, M.T. Zuber, L.N. Moresi, and M. Gurnis. "The role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection". In: *J. Geophys. Res.: Solid Earth* 105.B5 (2000), pp. 11, 063–11, 082. DOI: `10.1029/2000JB900003`.

[1524] S.J. Zhong, D.A. Yuen, L.N. Moresi, and M.G. Knepley. "7.05 - Numerical Methods for Mantle Convection". In: *Treatise on Geophysics (Second Edition)*. Ed. by Gerald Schubert. Second Edition. Oxford, 2015, pp. 197–222. DOI: `10.1016/B978-0-444-53802-4.00130-5`.

[1525] Shijie Zhong. "Analytic solutions for Stokes' flow with lateral variations in viscosity". In: *Geophy. J. Int.* 124.1 (1996), pp. 18–28. DOI: `10.1111/j.1365-246X.1996.tb06349.x`.

[1526] Shijie Zhong and Bradford H Hager. "Entrainment of a dense layer by thermal plumes". In: *Geophy. J. Int.* 154.3 (2003), pp. 666–676. DOI: 10.1046/j.1365-246X.2003.01988.x.

[1527] S-S Zhou and X-L Gao. "Solutions of the generalized half-plane and half-space Cerruti problems with surface effects". In: *Zeitschrift für angewandte Mathematik und Physik* 66.3 (2015), pp. 1125–1142. DOI: 10.1007/s00033-014-0419-4.

[1528] D.Y. Zhu, C.F. Lee, and K.T. Law. "Determination of bearing capacity of shallow foundations without using superposition approximation". In: *Can. Geotech. J.* 40 (2003), pp. 450–459. DOI: 10.1139/t02-105.

[1529] G. Zhu, T.V. Gerya, P.J. Tackley, and E. Kissling. "Four-dimensional numerical modeling of crustal growth at active continental margins". In: *J. Geophys. Res.: Solid Earth* 118 (2013), pp. 4682–4698. DOI: 10.1002/jgrb.50357.

[1530] Yongning Zhu and Robert Bridson. "Animating sand as a fluid". In: *ACM Transactions on Graphics (TOG)* 24.3 (2005), pp. 965–972. DOI: 10.1145/1073204.1073298.

[1531] O. Zienkiewicz and S. Nakazawa. "The penalty function method and its application to the numerical solution of boundary value problems". In: *The American Society of Mechanical Engineers* 51 (1982). DOI: xxxx.

[1532] O.C. Zienkiewicz. "Visco-plasticity, plasticity, creep and visco-plastic flow". In: *Lecture Notes in Mathematics* 461 (1975), pp. 297–328.

[1533] O.C. Zienkiewicz and I.C. Cormeau. "Visco-plasticity and creep in elastic solids - A unified numerical solution approach". In: *Int. J. Num. Meth. Eng.* 8 (1974), pp. 821–845. DOI: 10.1002/nme.1620080411.

[1534] O.C. Zienkiewicz and I.C. Cormeau. "Visco-plasticity and plasticity. An alternative for Finite Element solution of material nonlinearities". In: *Lecture notes in Computer Science* 10 (1974), pp. 259–287. DOI: 10.1007/BFb0015179.

[1535] O.C. Zienkiewicz and P.N. Godbole. "Flow of plastic and visco-plastic solids with special reference to extrusion and forming processes". In: *Int. J. Num. Meth. Eng.* 8 (1974), pp. 3–16. DOI: 10.1002/nme.1620080102.

[1536] O.C. Zienkiewicz, M. Huang, and M. Pastor. "Localization problems in plasticity using Finite Elements with adaptive remeshing". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 19 (1995), pp. 127–148. DOI: 10.1002/nag.1610190205.

[1537] O.C. Zienkiewicz, C. Humpheson, and R.W. Lewis. "Associated and non-associated visco-plasticity and plasticity in soil mechanics". In: *Géotechnique* 25.4 (1975), pp. 671–689. DOI: 10.1680/geot.1975.25.4.671.

[1538] O.C. Zienkiewicz, P.C. Jain, and E. Oñate. "Flow of solids during forming and extrusion: some aspects of numerical solutions". In: *Int. J. Solids Structures* 14 (1978), pp. 15–38. DOI: 10.1016/0020-7683(78)90062-8.

[1539] O.C. Zienkiewicz, M. Pastor, and Maosong Huang. "Softening, localisation and adaptive remeshing. Capture of discontinuous solutions". In: *Computational Mechanics* 17 (1995), pp. 98–106. DOI: 10.1007/BF00356482.

[1540] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method. Vol. 1: The basis.* Butterworth and Heinemann, 2002. ISBN: 0-7506-5049-4.

[1541] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method. Vol. 2: Solid Mechanics.* Butterworth and Heinemann, 2002.

[1542] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method. Vol. 3: Fluid Dynamics.* Butterworth and Heinemann, 2002.

[1543] O.C. Zienkiewicz, R.L. Taylor, and D.D. Fox. *The Finite Element Method for solid and structural mechanics*. Elsevier B.H., 2014.

[1544] O.C. Zienkiewicz, J.P. Vilotte, and S. Toyoshima. "Iterative method for constrained and mixed approximation. An inexpensive improvement of FEM performance". In: *Computer Methods in Applied Mechanics and Engineering* 51 (1985), pp. 3–29. DOI: 10.1016/0045-7825(85)90025-8.

[1545] O.C. Zienkiewicz and J. Wu. "Incompressibility without tears - How to avoid restrictions of mixed formulations". In: *Int. J. Num. Meth. Eng.* 32 (1991), pp. 1189–1203. DOI: 10.1002/nme.1620320603.

[1546] F. Zinani and S. Frey. "Galerkin Least-Squares Solutions for Purely Viscous Flows of Shear-Thinning Fluids and Regularized Yield Stress Fluids". In: *J. of the Braz. Soc. of Mech. Sci. & Eng.* XXIX (2007), pp. 432–443.

[1547] S. Zlotnik, P. Diez, M. Fernandez, and J. Verges. "Numerical modelling of tectonic plates subduction using X-FEM". In: *Computer Methods in Applied Mechanics and Engineering* 196 (2007), pp. 4283–4293. DOI: 10.1016/j.cma.2007.04.006.

[1548] S. Zlotnik, M. Fernandez, P. Diez, and J. Verges. "Modelling gravitational instabilities: slab break-off and Rayleigh-Taylor diapirism". In: *Pure Appl. Geophys.* 165 (2008), pp. 1491–1510. DOI: 10.1007/s00024-004-0386-9.

[1549] M.T. Zuber and E.M. Parmentier. "Lithospheric necking: a dynamic model for rift morphology". In: *Earth Planet. Sci. Lett.* 77 (1986), pp. 373–383. DOI: 10.1016/0012-821X(86)90147-0.

[1550] M.T. Zuber, E.M. Parmentier, and R.C. Fletcher. "Extension of Continental Lithosphere: A Model for Two Scales of Basin and Range Deformation". In: *J. Geophys. Res.: Solid Earth* 91.B5 (1986), pp. 4826–4838. DOI: 10.1029/JB091iB05p04826.

[1551] T. Zwinger, R. Greve, O. Gagliardini, T. Shiraiwa, and M. Lyly. "A full Stokes-flow thermo-mechanical model for firn and ice applied to the Gorshkov crater glacier, Kamchatka". In: *Annals of Glaciology* 45 (2007), pp. 29–37.

General index