ICML 2022

International Conference On Machine Learning

Born-Infeld (BI) for AI: Energy-Conserving Descent (ECD) for Optimization + Sampling

arXiv:2201.11137 [cs.LG]

with G. Bruno De Luca

Offshoot of discussions w/ J. Batson, Y. Kahn, D. Roberts on inflation and optimization; +early/intermediate collaboration with G. Panagopoulos, Thomas Bachlechner

+New discussions/collaborations: ML (Kunin), Quantum Chemistry (Zhang), Sampling (Robnik/Seljak; Cheng)

Verlinde Symposium 2022

Happy 60th!

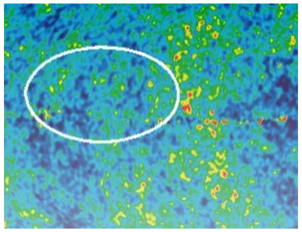
Twin pillars of theoretical physics

Many thanks for all the brilliant contributions propelling the field forward!

My own research owes much to Erik's and Herman's works for decades,
recent synergy in solvable versions of (A)dS patch-wise holography.

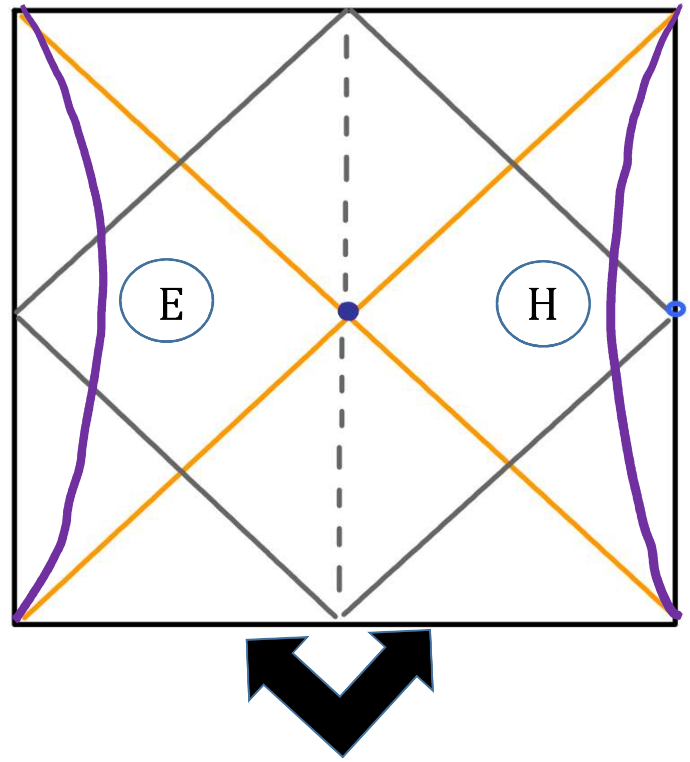
Herman and Erik's incisive creativity inspire and enrich the entire community.

Indeed, much like S. Hawking,   Erik and Herman are embedded in the universe:



- `de Sitter Holography with a finite number of states'
- `Holography and Compactification'
- 'Moving the CFT into the bulk with $T\bar{T}$'
- `Quasi-local energy and microcanonical entropy in two-dimensional nearly de Sitter gravity''
- …

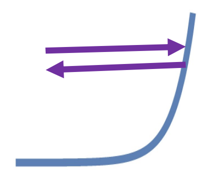
- 'de Sitter Holography and entanglement entropy', dS/dS and $T\bar{T}$, $T\bar{T}$ and EE, de Sitter microstates from the $T\bar{T} + \Lambda_2$ deformation and the Hawking/Page transition' w/Coleman, Dong, Gorbenko, Lewkowycz, Liu, Mazenc, Soni, Shyam, Torroba, Yang,…



Twin Highly Mixed Sectors

- **`Matrix string theory'**:

Question: Do Dirichlet walls exist in string/M theory?   Generalized Liouville wall? If so, Non-fluctuating timelike boundaries possible in more general spacetimes than AdS



$$\mathit{Matrix\ Theory\ Hamiltonian} = \sum Tr(\dot{X}^2) + Tr[X^M, X^N]^2 + \mathrm{Tr}\ O_\kappa \exp(\kappa X^{(10)})$$

$$\text{String theory worldsheet action} = \mathit{tension} * \int (G_{MN}\partial X^M \partial X^N + O_\kappa \exp(\kappa X^{(9)}))$$

Exec summary: *real dressed spectrum* of the universal and solvable

$T\bar{T} + \Lambda_2$ deformation

$$\frac{\partial}{\partial\lambda} \log Z = -2\pi \int d^2x\,\sqrt{g}\,\langle T\bar{T}\rangle + \frac{1-\eta}{2\pi\lambda^2}\int d^2x\,\sqrt{g}$$

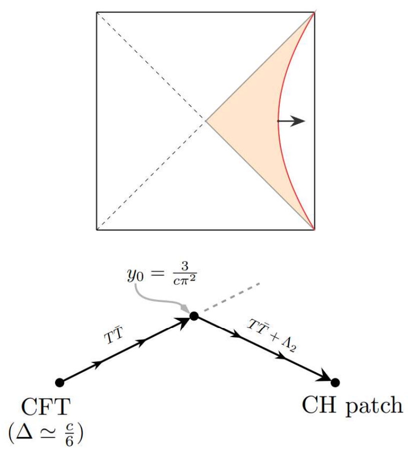
Zamalodchikov et al, Dubovsky et al, Cavaglia et al … Gorbenko ES Torroba '18

of a CFT on a cylinder captures (only) the microstates and the geometry of the dS_3 observer patch Shyam, Coleman et al '21

$$\mathcal{E} = \frac{1}{\pi y}\left(1 \mp \sqrt{\eta + \frac{y}{y_0}(1-\eta) - 4\pi^2 y\left(\Delta - \frac{c}{12}\right) + 4\pi^4 y^2 J^2}\right)$$

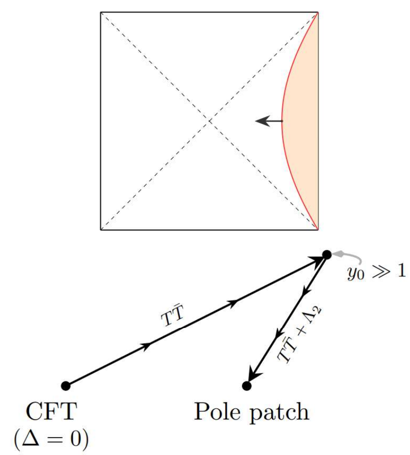
Cosmic horizon patch

(Dressed $\Delta \simeq \frac{c}{6}$ black hole microstates)



Pole patch

(Dressed $\Delta = 0$ vacuum)



$\mathcal{E} = \frac{1}{\pi y}\left(1 + \sqrt{\eta + \ldots}\right)$ $\longleftarrow$ related by $\pm\sqrt{\ }$ $\longrightarrow$ $\mathcal{E} = \frac{1}{\pi y}\left(1 - \sqrt{\eta + \ldots}\right)$

BPS black hole state counting (Strominger/Vafa…), used extended SUSY to control weak → strong coupling deformations preserving state count. Here we have a **new type of controlled deformation** applicable to dS, again preserving state count: 'integrable deformation' of non-integrable seed theory.

ICML 2022

Born-Infeld (BI) for AI: Energy-Conserving Descent (ECD) for Optimization + Sampling

arXiv:2201.11137 [cs.LG]

with G. Bruno De Luca

Offshoot of discussions w/ J. Batson, Y. Kahn, D. Roberts on inflation and optimization; +early/intermediate collaboration with G. Panagopoulos, Thomas Bachlechner

+New discussions/collaborations: ML (Kunin), Quantum Chemistry (Zhang), Sampling (Robnik/Seljak; Cheng)

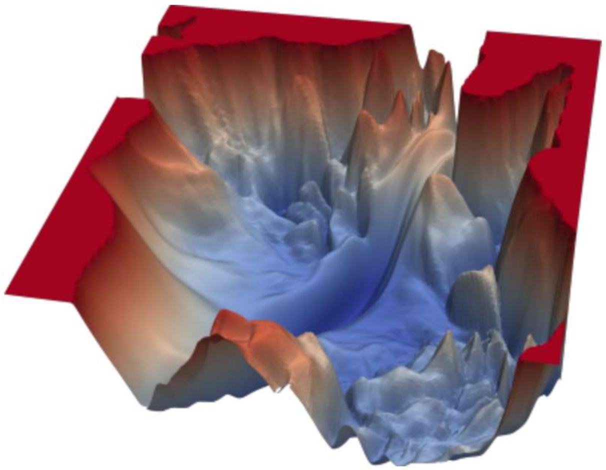
Verlinde Symposium 2022

- For many problems one needs to minimize an objective (`Loss') function V, descending a generally non-convex high dimensional landscape.

  --data analysis/machine learning

  -- PDE solving, $\text{Loss} = \sum (PDEs)^2 + (boundary\ conditions)^2$: want global min

Gradient descent methods and variants can work well w/modern tweaks, but sometimes get stuck and/or don't sample all desired solutions.



Early U cosmology: models for descending a potential landscape V.
--Example: DBI: relativistic speed limit $\rightarrow 0$ as $V \rightarrow 0$ *without friction, consistent with energy conservation* → *calculability*

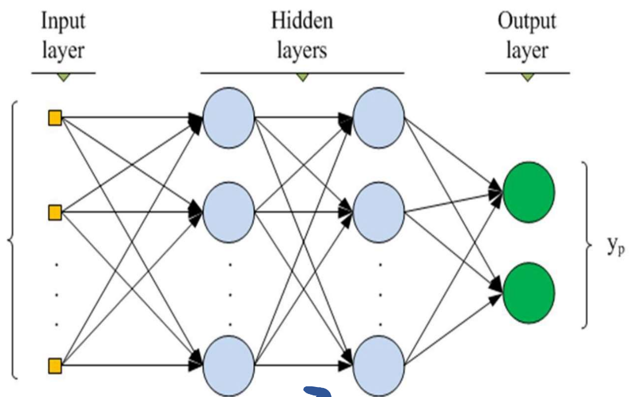
cf Relativistic Gradient Descent Franca et al '19 (with constant speed limit)

- Another common goal is Sampling from a distribution, e.g. multimodal.

Schematic of NN's for ML & PDE solving

Lagaris, Likas, Fotiadis '97,..., Douglas, Lakchminarasimhan, Qi '20, ...

Points sampled from domain of PDE



Output functions (ansatzes for functions being solved for)

$\sigma(w \cdot x + b)$

denote as $\vec{\theta}$ (NN parameters)

Repeated application builds up nonlinear output functions/ansatzes

Then form loss functional: e.g.

$$\sum_{pts,eqs} (PDEs)^2 + (boundary\ conditions)^2$$

Descend the loss landscape via gradient descent or generalizations

Also ML beyond classical PDEs: QFTs

P. de Haan, C. Rainone, M.C.N. Cheng and R. Bondesan, *Scaling Up Machine Learning For Quantum Field Theory with Equivariant Continuous Flows*, 2110.02673.

J. Halverson, *Building Quantum Field Theories Out of Neurons*, 2112.04527.

(Supervised) Machine Learning: e.g. $Loss \sim \sum_{\{x\}} (y_{output} - y_{true})^2$

Quantum Chemistry: $Loss \sim \langle\, H \,\rangle_{trial\ wavefunction}$

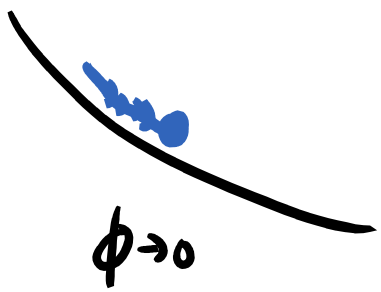
Early Universe inflation requires nearly constant potential $V(\phi)$

- Slow roll (flat potential, Hubble friction dominates)
- Interactions slow the field, e.g. DBI inflation: speed limit ϕ-dependent

$$S = -\int d^4x \left\{ \frac{\phi^4}{\lambda}\sqrt{1-\frac{\lambda\dot{\phi}^2}{\phi^4}} + \Delta V(\phi) \right\}$$

Testable (falsifiable(?)) via non-Gaussianity ($\simeq$equilateral shape)

Alishahiha, ES, Tong '04



$f_{NL}^{DBI} = 14 \pm 38$ Planck

$f_{NL}^{local} = -0.9 \pm 5.1$; $f_{NL}^{equil} = -26 \pm 47$; and $f_{NL}^{ortho} = -38 \pm 24$ (68 % CL, statistical)

$\boxed{\text{Distinct behavior and predictions from slow roll}}$

Non-gravitational version conserves energy (no friction), only stopping at V=0

$$S = -\int V(\vec{\theta})\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V(\vec{\theta})}}$$

$$\pi_i = \frac{\partial L}{\partial \dot{\theta}^i} = \frac{\dot{\theta}_i}{\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V}}}$$

$$H = \frac{V}{\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V}}} = \sqrt{V(V + \vec{\pi}^2)} \equiv E = constant$$

$\Rightarrow$ **Cannot** stop at local min, even without stochastic noise (but can get stuck in orbit).
Cannot overshoot V=0.
Faster in shallow valleys.

Distinct behavior from gradient descent

Phase space volume strongly dominated near global minimum:

$$Vol(\mathcal{M}) = \frac{2\pi^{n/2}}{\Gamma(n/2)}\int d^n\theta \int d\tilde{\pi}\tilde{\pi}^{n-1}\delta(\sqrt{V(V+\tilde{\pi}^2)} - E) = \frac{2\pi^{n/2}}{\Gamma(n/2)}\int d^n\theta \frac{E}{V}\left(\frac{E^2}{V} - V\right)^{\frac{n-2}{2}}$$

Many variations on this theme, e.g. 2-derivative action with mass ~ 1/Loss

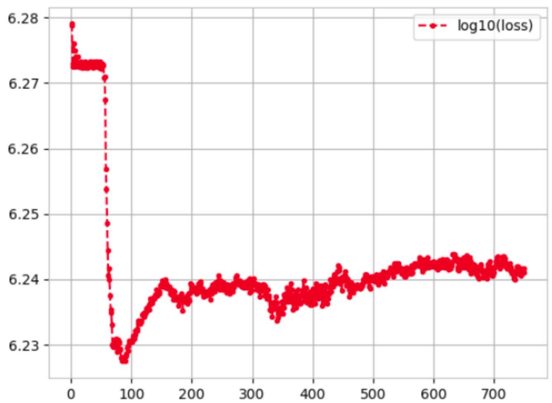
Energy Conserving Descent (ECD)

As an energy conserving dynamical system in a rich loss landscape (without symmetries), BI can easily be chaotic, with random initialization avoiding stable orbits.

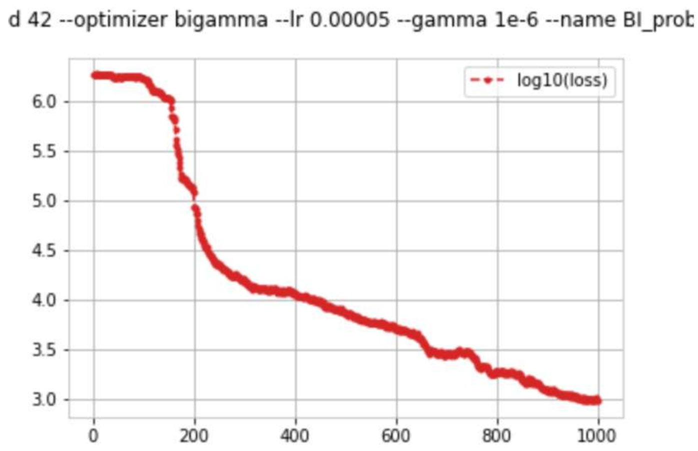
But if a particular problem (NN & Loss function) leads to long-lived orbits, we can add extra features to the algorithm (as in chaotic billiards problems) to stimulate faster mixing

Toy Example: $-\nabla^2 u + u^2 = f$, $f = \frac{1}{8}\left(3 - 4(1 + 6400(x_1^2 + x_2^2))\cos(40(x_1^2 + x_2^2)) + \cos(80(x_1^2 + x_2^2)) - 640\sin(40(x_1^2 + x_2^2))\right)$

Original problem (stuck in orbit):



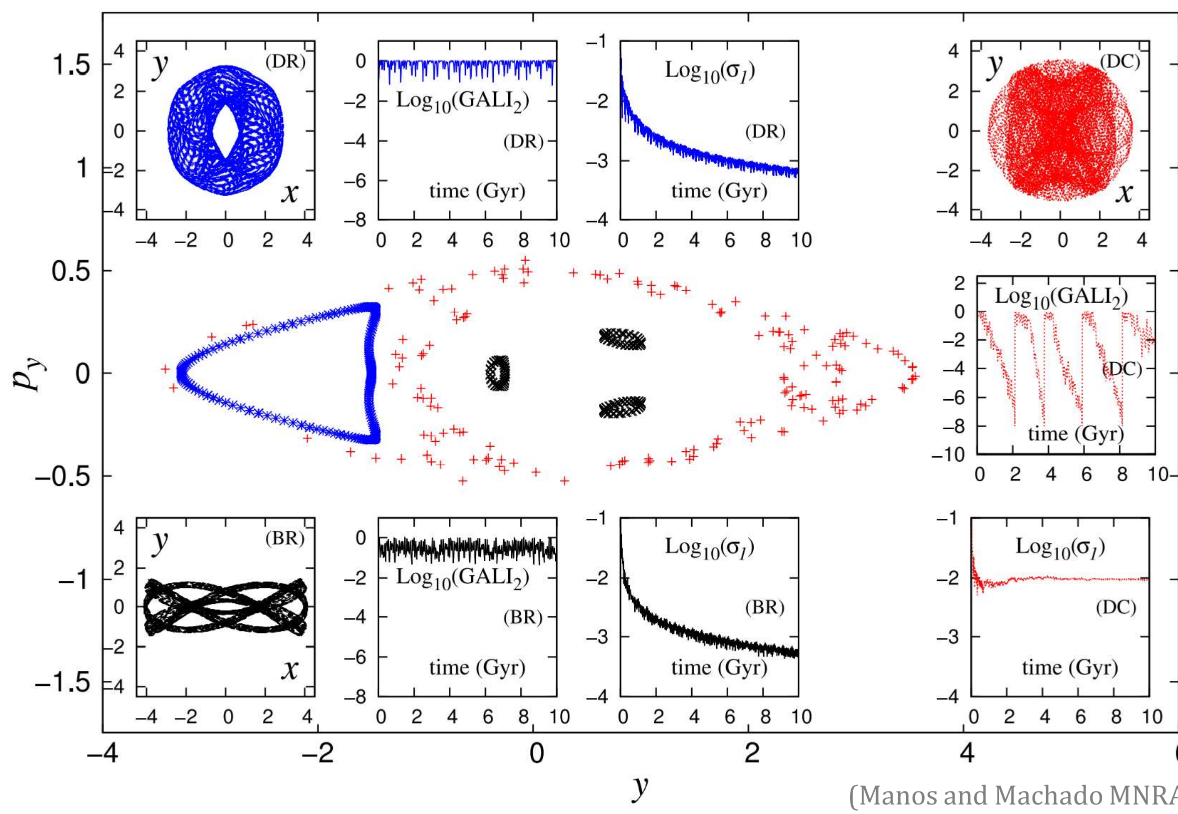
With added feature (unstuck):



Our redshifted BI dynamics is a bit like galactic dynamics, solar system, … where chaos (as well as long lived orbits) is familiar.

We add elements aimed at ensuring rapid mixing.

Chaos and dynamical trends in barred galaxies 2209



(Manos and Machado MNRAS '14).

Figure 5. The Poincaré Surface of Section defined by $x = 0$, $p_x \geq 0$ with $H = -0.19$, for three typical orbits (two regular and one chaotic) being integrated for 10 Gyr. The set of parameters for the bar, disc and halo components are chosen from the fits with the 3-d.o.f. TD Hamiltonian at $t = 7.0$ Gyr of the N-body simulation. In the insets, we depict their projection on the (x, y)-plane together with the GALI_2 and MLE σ_1 evolution in time (see Table 1 for the exact parameters and text for more details on these trajectories).

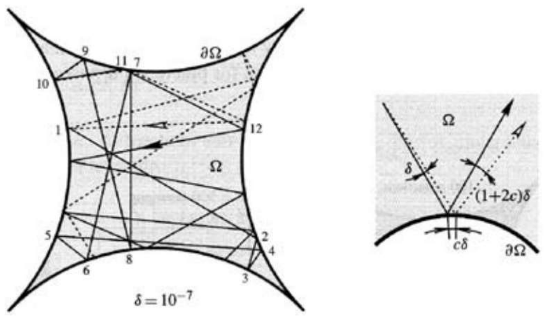
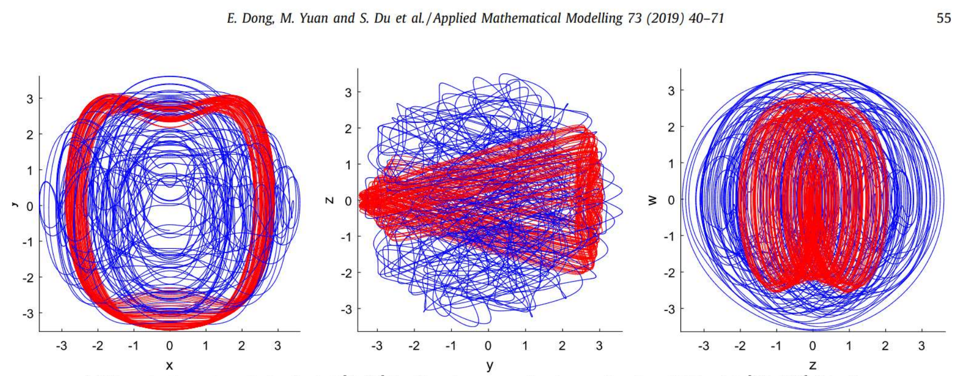


Figure 2. Illustration of the trajectory sensitivity to the initial conditions in a billiard model with convex borders.



(a) Transient quasi-periodic for $t \in [0, 50]$ (red) and conservative hyperchaotic orbit for $t \in [50, 100]$ (blue);

Adding dispersing elements, (e.g. billiards or negative curvature) supports mixing (decay of correlations)

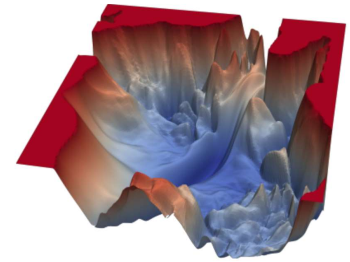
After some time, for a particle p in a droplet and phase space region R,

$$Prob(p \in R) \propto Vol(R)$$

(>ergodicity: $\langle f \rangle_t = \langle f \rangle_{phase\ space}$)

Overview: Optimization of an objective function F

- Data analysis/Machine Learning [F = loss]
- Solving (Partial) Differential Equations
  [F = Σ (PDEs)^2+(boundary conditions)^2]
- Many scientific applications



[Image from Li et al. , '18]

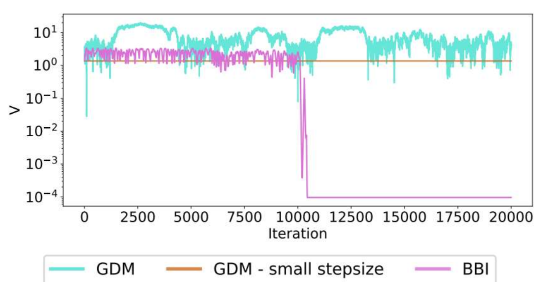
Gradient Descent with Momentum (GDM) can work well with modern tweaks.

Physical analogue: particle motion on potential energy V = F, *with friction*, discretized.

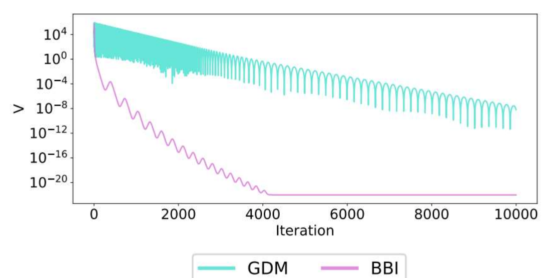
Our proposal: Energy Conserving Descent (**ECD**): discretized physical evolution, *without friction*, nonetheless slowing near minimal F. Examples include:

- BBI: relativistic, (speed limit)^2 = V = F-ΔV [or more general (speed limit)^2= g(V)]
- Ruthless: non-relativistic, mass ∝ 1/g(V)

Ackley 2d (nonconvex)



Zakharov 10d (shallow)



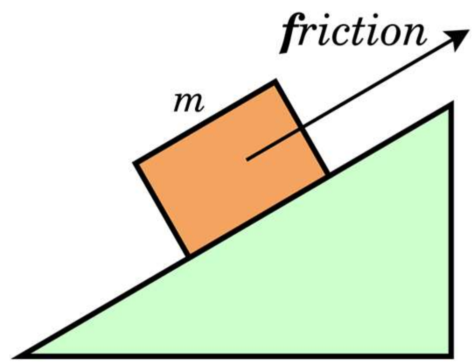
+ other synthetics, PDEs, small ML (Cifar, MNIST, Tiny ImageNet [new]), chemistry, sampling [new]

No friction ⟹ **Energy Conservation** ⟹ favorable properties and improved calculability: concrete formula for distribution of results: in all dims weighted toward small V = F-ΔV

Physics of GDM

Particle descending a potential energy landscape V

$$V(\Theta) = F(\Theta) - \Delta V$$



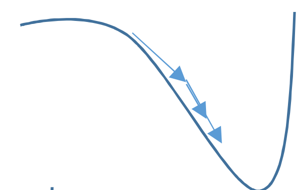
Familiar law of motion: Force = mass × acceleration

$$-\nabla V - \boldsymbol{f}\dot{\Theta} = m\,\ddot{\Theta}$$

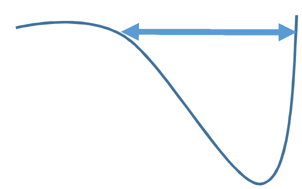
Friction coefficient $\boldsymbol{f}$ $\Longrightarrow$ Energy not conserved

First-order form: $p = m\dot{\Theta}$ $\quad \dot{p} = -p\dfrac{\boldsymbol{f}}{m} - \nabla V$

Discretization → GD with Momentum (GDM) + minibatches → SGDM



- Energy $E = \dfrac{p^2}{2m} + V(\Theta)$ not conserved because of friction



- $\boldsymbol{f} = 0$ would conserve energy, but the particle flies quickly past V≃0, spending very little time there (especially in high dimensions)

ECD: physical dynamics can conserve energy yet slow near V=0

Next: explicit realizations

Explicit realizations of ECD

Change the dynamics to conserve Energy E and favor V≃ 0

General structure:

$$H(\Theta, \Pi) = E$$

Position vector (parameters)

Momentum vector

Dynamical equations (cf. Newton's laws of motion): $\dot{\Theta} = \frac{\partial H}{\partial \Pi}, \qquad \dot{\Pi} = -\frac{\partial H}{\partial \Theta}$

1. **BI:** (speed limit)^2 = V = F-ΔV, [or general function g(V)]

$$H = \sqrt{g(V)\big(\Pi^2 + g(V)\big)} = g(V)/\sqrt{1 - \frac{\dot{\Theta}^2}{g(V)}}$$

- Cannot exceed relativistic speed limit: $\dot{\Theta}^2 \leq g(V)$ [ES, Tong, +Alishahiha '04, cf. França et al. '20]

2. **Rootless (Ruthless):** mass ∝ 1/g(V) $H = \left(\frac{\Pi^2}{2m(V)}\right) = \text{g(V)}\ \Pi^2 = \frac{1}{2}m(V)\dot{\Theta}^2$

- Slows as the particle gets heavy: $m(V) \to \infty, \ \ g(V) \to 0 \Rightarrow \dot{\Theta}^2 \to 0$

Building ECD optimization algorithms

0. Choose the continuum dynamical system

1. Discretize the continuum equations of motion
   - e.g. BI with g(V) = V:

$$\sqrt{V(V+\vec{\pi}^2)} \equiv E$$

$$\pi_i(t+\Delta t) - \pi_i(t) = -\Delta t \frac{\partial_i V(\mathbf{\Theta}(t))}{2}\left(\frac{E}{V} + \frac{V}{E}\right)$$

$$\theta_i(t+\Delta t) - \theta_i(t) = \Delta t\, \pi_i(t+\Delta t)\frac{V(\mathbf{\Theta}(t))}{E}$$

2. Choose an initialization
   - Common choice: Π(0) => E = V(0)
   - Option: E > 0 => choice of Π(0) compatible with Energy eq.

3. Use discretized equation as update rules

4. Add other features
   - Enforce strict Energy conservation rescaling Π
   - Adaptive tuning of shift ΔV = F-V (next page)
   - Option: random rotation of momenta ("bouncing", explained later)

5. Test it!

DATA SET	SGD	BBI
MNIST	99.166 , 98.160	99.177 , 99.190
CIFAR-10	92.628 , 92.655	92.434 , 92.435

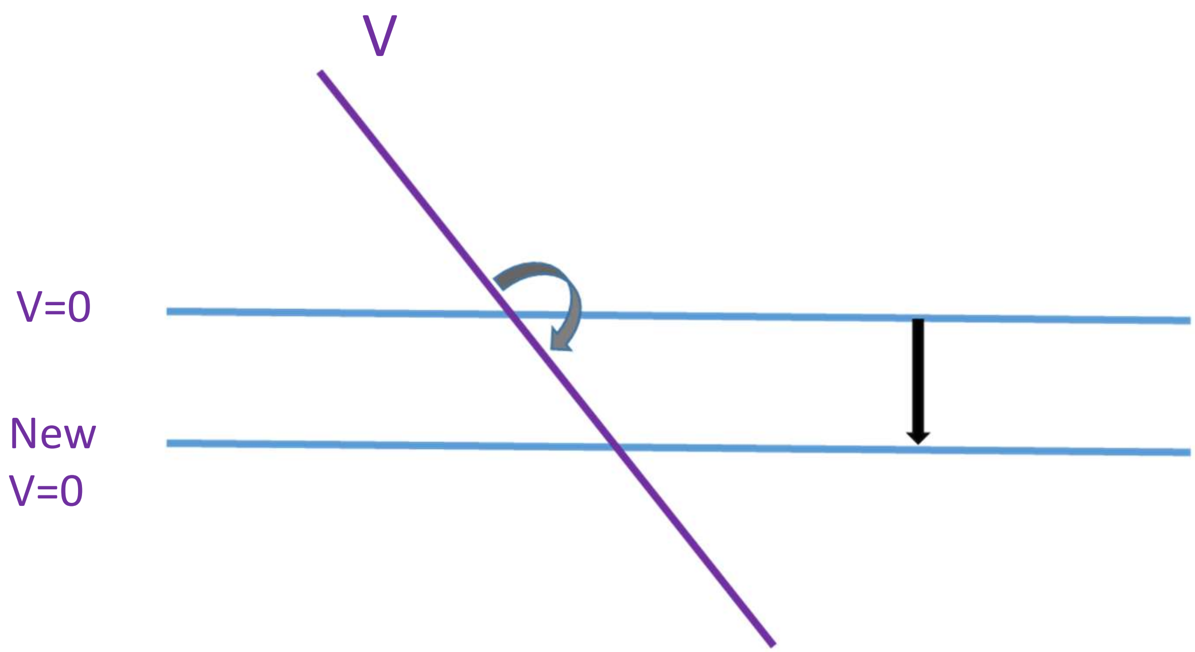
Modest (~50) statistics and limited hyper-parameter tuning (without all the tweaks on either side); just a check of basic competence. "Bouncing" not required here.

Automatic (adaptive) Tuning of ΔV

The value of the loss function F at the objective is not always known:

$$\boxed{V = F - \Delta V}$$

ΔV is a hyperparameter that can automatically adjust (recover from an over-estimate). New upgrade to optimizer code.



Given a too-high initial guess for ΔV, the loss extends to $V = F - \Delta V < 0$ and the trajectory will jump to a small negative value $V < 0$ due to the discreteness. Conditioned on this, ΔV may be lowered, iteratively tuning it.

Recap so far:

- Optimization of an objective function F
- Descent dynamics as (discrete) physical evolution on a potential V = F-ΔV
- Equations of motion (update rules) obtained from a Hamiltonian H
  - Gradient Descent with Momentum: a time-dependent H(Π, Θ, t)
    - Energy not conserved: $\dot{E} = -f\frac{\Pi^2}{m^2} \leqslant 0$
    - Simply removing friction (f =0) does not converge
- Alternative physics: Energy Conserving dynamical systems converging to V→0

$$E = H_{\rm ECD}(\Theta, \Pi)$$

  - Energy is conserved: $\dot{E} = 0$
  - 2 explicit examples: **BI** [relativistic], **Ruthless** [m = 1/g(V)].
- Discretization gives update rules → new optimization algorithms

Simple benchmarks show that the idea works: friction not needed for optimization.

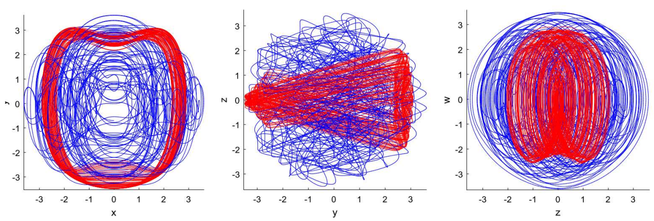
Next: advantages of conserving energy

Energy Conservation

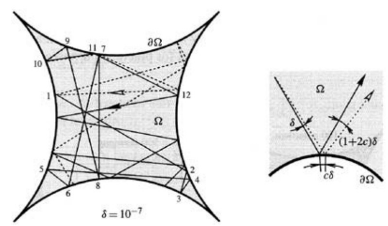
$$H = \frac{V}{\sqrt{1-\frac{\dot{\vec{\theta}}^2}{V}}} = \sqrt{V(V+\vec{\pi}^2)} \equiv E = constant \implies$$

- **Cannot** stop unless V=E or V=0, so cannot stop in high local minimum

Can get stuck in orbit at high V. Generically such orbits are unstable: chaos – sensitive dependence on initial conditions – is typical in physical systems. Nearby trajectories disperse roughly on a *mixing* timescale.



[Image from Dong, Yuan, Du et al. '19]



[Image from Encyclopedia of Nonlinear Science, '04]

Chaos and *mixing* has been **proven** in mathematical billiards problems.

This inspires optional Bounces in BI algorithm above to reduce the mixing time $\implies$ **BBI**

- *Phase space* (positions & momenta) *volume* is preserved under the evolution. $Vol(phase\ space) = \int d^n\Theta d^n\Pi\delta(H(\Pi,\Theta) - E)$

- Past the mixing time, the probability to find a particle from a droplet (bundle of trajectories) in a region M of phase space is ∝ Vol(M)

For ECD, phase space volume is strongly dominated near V=0:

$$Vol(\mathcal{M}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int d^n\theta \int d\tilde{\pi}\tilde{\pi}^{n-1}\delta(\sqrt{V(V+\tilde{\pi}^2)} - E) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int d^n\theta \frac{E}{V}\left(\frac{E^2}{V} - V\right)^{\frac{n-2}{2}}$$

For $V \to 0$, $\quad Vol \propto \int \frac{d^n\Theta}{V^{n/2}} = \int d\Omega \int d|\Theta||\Theta|^{n-1}\frac{1}{V^{n/2}}$

For a basin V $\sim |\Theta|^2$, this becomes $\sim \int d\Omega \int d|\Theta|/|\Theta|$

$V \to g(V) \sim V^\eta \quad \eta > 1$ enhances the preference for V=0 (beats the effect of high dimension n!)

(g(V) also useful for sampling, in addition to optimization)

[GBDL, Roblik, Seljak, ES in progress]

- In contrast, pure momentum would not favor small V:

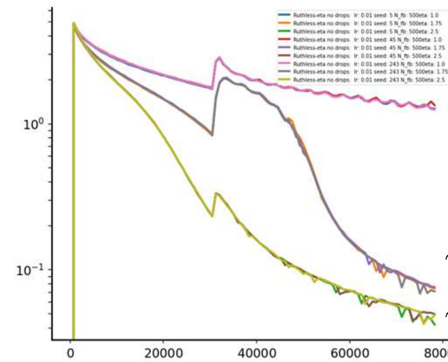
$$\mathrm{Vol}(\mathcal{M}) \propto \int d^n\theta (E-V)^{\frac{n-2}{2}} \quad \text{frictionless non-relativistic momentum}$$

- The volume formula would not apply at all with friction (less predictive in that sense).

Exploiting the volume formula for image classification (preliminary)

- Enhancement of volume density for η>1 near a quadratic minimum V ~ θ^2:

$$\mathrm{vol} \propto |\Theta|^{n(1-\eta)-1} d|\Theta|$$



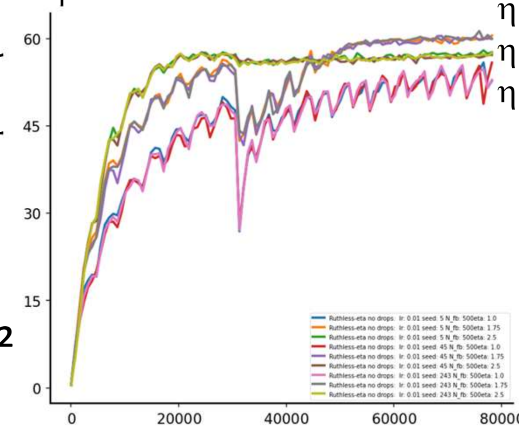
- Small Tests on Tiny-ImageNet* with D. Kunin
  (+ImageNet 1K in progress)

Protocol: lr = 0.01, **no** lr drop needed, 500 bounces,

Averaging of late-epoch weights (SWA) [Izmailov et al. '19]

Training loss decreases monotonically with η, improving test accuracy for intermediate η>1

m=1/V^η	Accuracy	Accuracy (weights averaged)
η=1	55.44	62.12
η=1.75	61.3	64.1



Compared with SGD: with lr drops

(start 0.1, drop factor 0.1@ep. [30,60,80]) :

Accuracy: 62.52, Accuracy (weights averaged): 62.93

SGD: without lr drops is worse, as well as with loss → loss²

[ECD also > best comparable SGDM in cf. Li et al. '21, Tanaka, Kunin et al. '20...]

*ResNet-18, epochs: 100, batch size: 128, weight decay: 10^{-4}, loss: Cross Entropy

Testing the volume formula

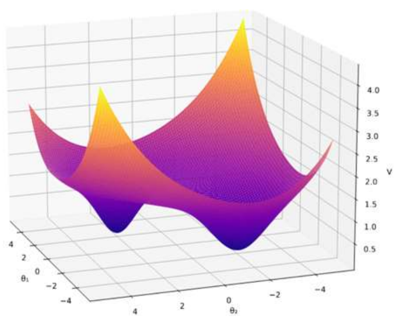
Evaluated in different regions predicts distribution of results (given mixing)

For g(V) = V: $$Vol(\mathcal{M}_{\mathcal{I}}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} E^{n-1} \int d^n(\theta - \theta_I) V^{-n/2}$$

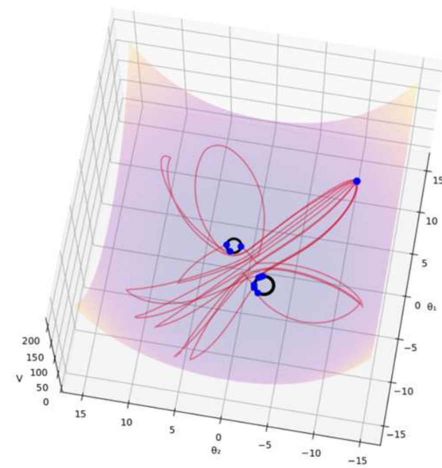
Near a minimum:

$$V \simeq V_I + \frac{1}{2}\sum_{i=1}^{n} m_{Ii}^2 (\theta_i - \theta_{Ii})^2 \qquad Vol(\mathcal{M}_{\mathcal{I}}) \to b_n \left(\frac{2\pi^{n/2}}{\Gamma(n/2)}\right)^2 \frac{E^{n-1}}{\prod_i m_{Ii}} \log(V_I) \qquad V_I \to 0$$

Empirical check:

V = 

Bouncing trajectories find the 2 basins:



Prediction of ratio of convergence:

$$\frac{\mathrm{Vol}(\mathcal{M}_1)}{\mathrm{Vol}(\mathcal{M}_2)} \sim 1.93$$

Results:

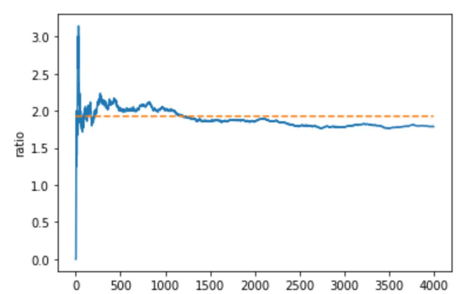


Figure 4: Partial ratios.

Agreement within 10%

Behavior in shallow regions

Volume formula prefers flatter minima

ML lore: flatter minima generalize better

$$V \simeq V_I + \frac{1}{2}\sum_{i=1}^{n} m_{Ii}^2(\theta_i - \theta_{Ii})^2$$

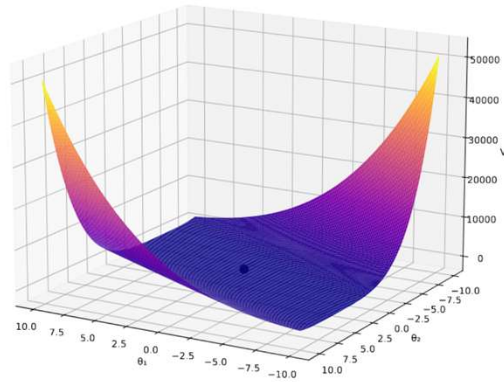
$$Vol(\mathcal{M}_\mathcal{I}) \to b_n \left(\frac{2\pi^{n/2}}{\Gamma(n/2)}\right)^2 \frac{E^{n-1}}{\prod_i m_{Ii}} \log(V_I) \quad V_I \to 0, \qquad m_{iI}^2 \longrightarrow 0$$

Prediction: BI is faster on shallow directions than GD

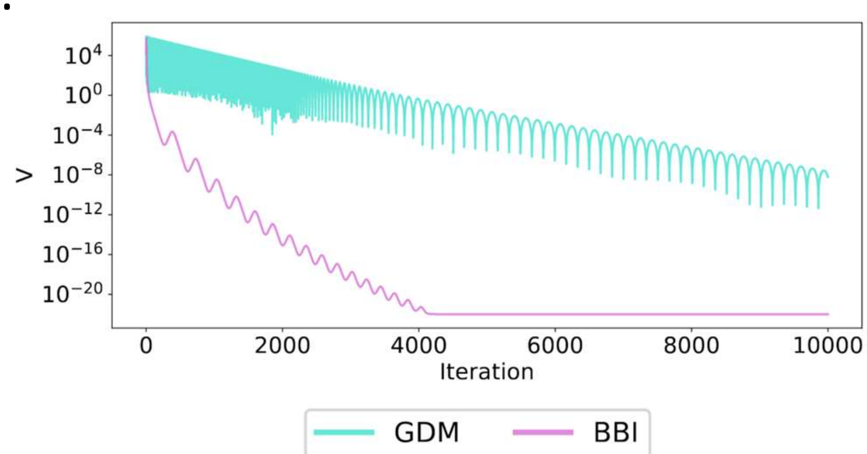
$$\Theta \sim e^{-mt/\sqrt{2}} \quad \text{vs} \quad \Theta \sim e^{-m^2 t/f}$$

Empirical check:

V = 10-dimensional Zakharov function



Results:



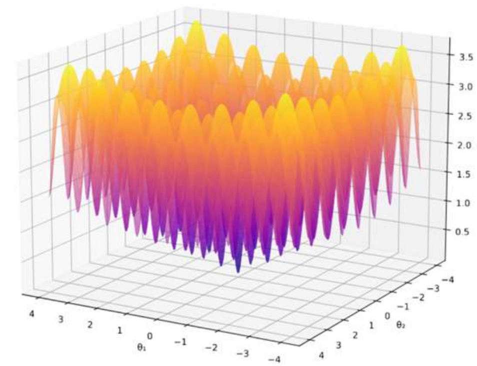
Hyperparameters tuned with hyperopt

Avoiding high local minima

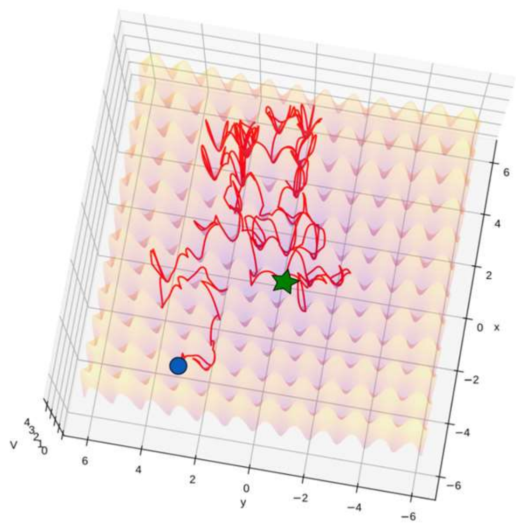
Energy conservation: **ECD** cannot stop in high local minima

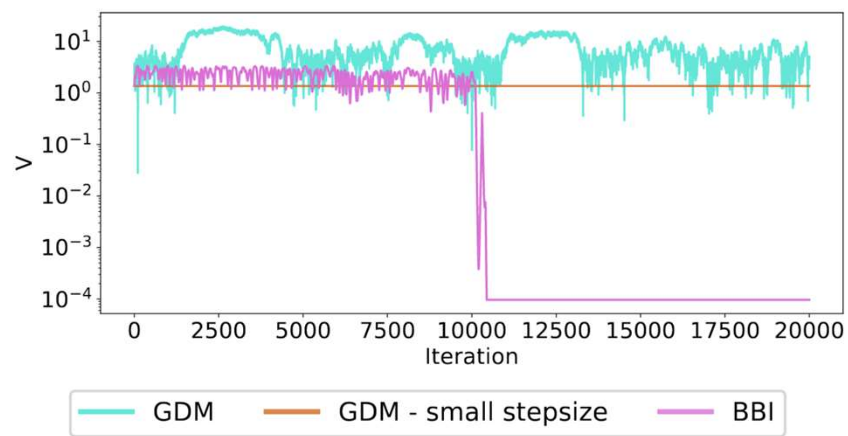
Empirical check: Highly non-convex function

V = 2-dim Ackley function :



Results:





BBI explores and finds the global minimum

Hyperoptimized fixed lr, and for GDM also momentum. GDM either stuck in initial basin or helped out by `catapult' mechanism [Lewkowycz et al. '20], , then more erratic (not settling in global minimum).

Summary comparison

ECD	Friction ((S)GDM, ...)
Conserves energy E	Friction drains E
Cannot get stuck in high local minimum	Can stop in high local minimum
Cannot overshoot $V = 0 = \nabla V$	Can overshoot $V = 0 = \nabla V$
Depends on V and ∇V	depends only on ∇V
On shallow region: $\theta \sim e^{-mt/\sqrt{2}}$	On shallow region: $\theta \sim e^{-m^2 t/f}$
Analytic prediction for distribution	Stochastic intuition for distribution
Generalizes	Generalizes

Generalization ok:
speed limit kicks in for
V ≪ E, Vol(phase space)
favors flat basins.

Statements persist with noise (mini-batches) in our prescription:
BBI speed limit tamps down noise, while the bounces (when needed) provide controlled stochasticity for short mixing time.

Application: Solving Partial Differential Equations

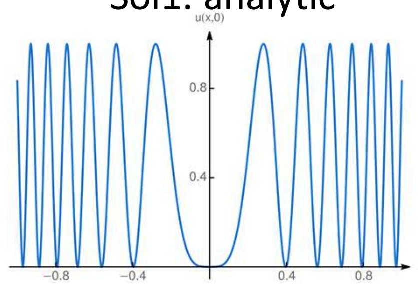
- Most common strategy with ML tools: a NN as ansatz for the PDE: [Lagaris et al. '98, ..., Raissi et al. '19,..]

$$F = V = \sum_{x \in \text{domain}} \text{PDE}[\mathcal{N}(x;\Theta)]^2 + \gamma \sum_{x \in \text{boundary}} \text{BC}[\mathcal{N}(x;\Theta)]^2 + R(\Theta)$$

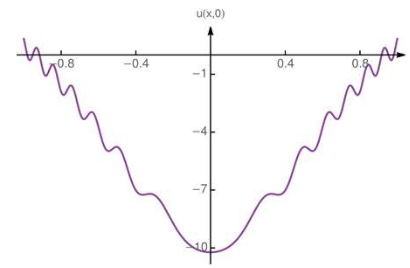
- We reverse-engineered hard (highly nonlinear) 2d PDEs with known multiple solution and checked if ECD optimization finds them

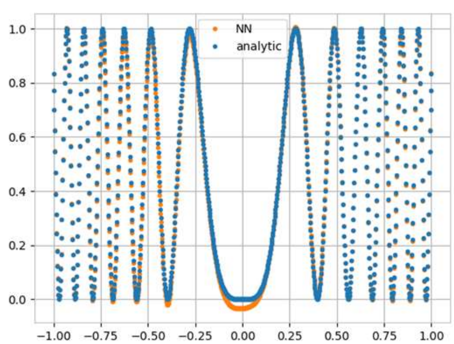
1d slices of *known* solutions:

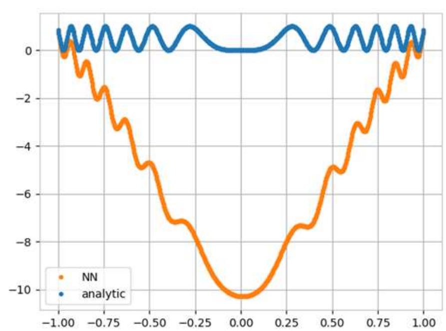
Sol1: analytic



Sol2: numerical







1d slices of *learned* solutions

Found **both** from same initialization: bounces distribute results (mixing)

Another common goal is **Sampling from a distribution**. To sample

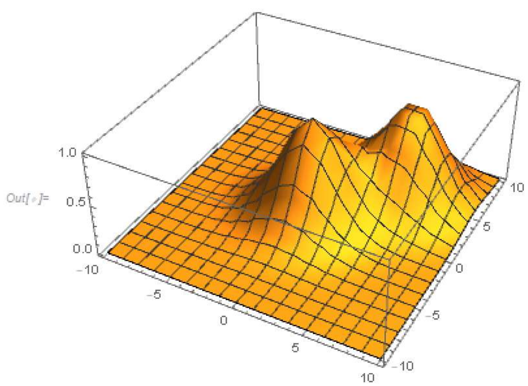
$$\exp(-F(\Theta))$$

using ECD, e.g. the version with Hamiltonian:

$$H = g(F)\Pi^2 = E$$

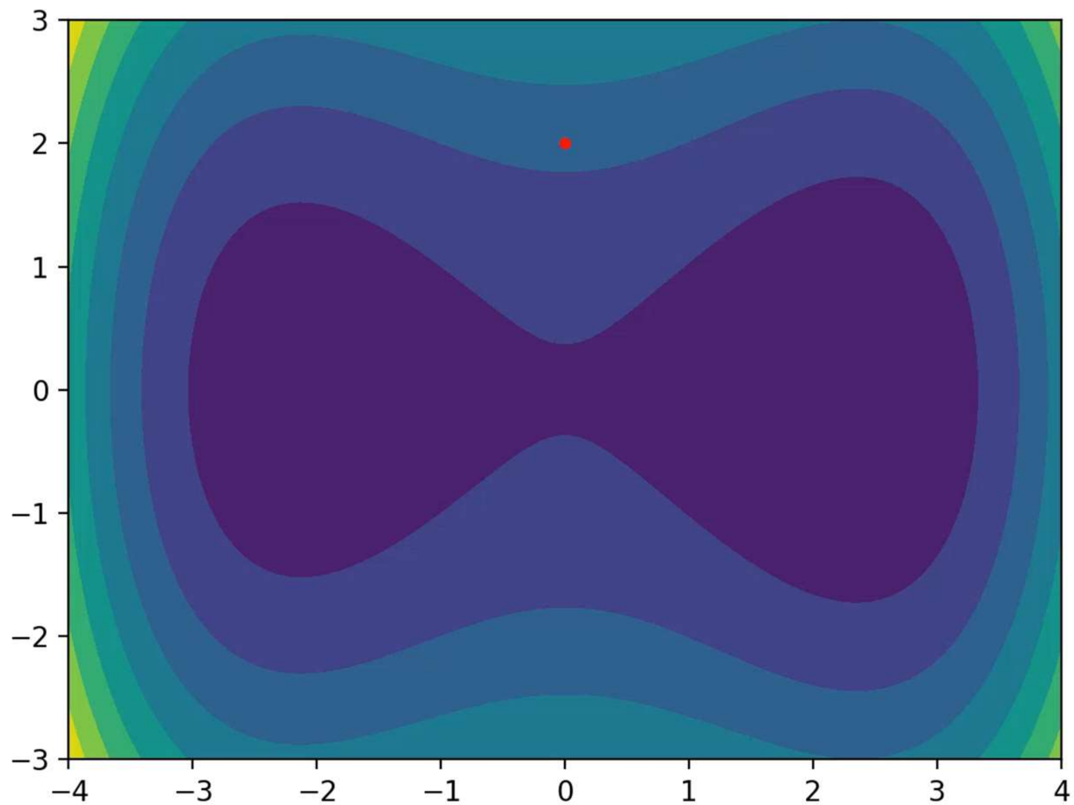
we again use the fixed-energy phase space volume formula:

$\int d^n\Pi \int d^n\Theta \exp(-\mathrm{F})\delta\big(E - H(\Theta,\Pi)\big) \propto \int d^n\Theta \exp(-F)$ requires $g(F) = \exp(2\frac{F}{n})$

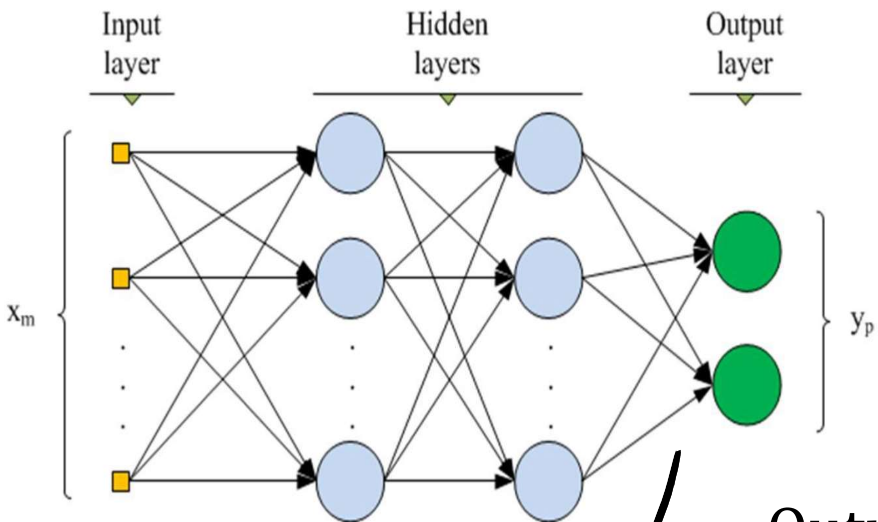


Reproduces distributions in warmups.
Will compare performance to
existing sampling methods,
e.g. Hamiltonian Monte Carlo (very
different). With Robnik, Seljak; Cheng

[Roblik/Seljak]



Feature/Representation Learning:



Output layer linear

$$y(x) = W^3 \underbrace{\sigma(W^2\sigma(W^1x + b^1)} + b^2)$$

Hidden layer updates required for feature learning (otherwise linear regression)

Std large-width limit → linear => feature learning ~ Depth/Width

Roberts, Yaida, Hanin

Other large-width limits preserve feature learning Yang/Hu

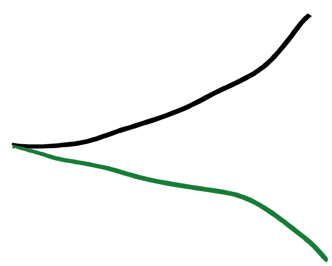
Optimizer-dependent…

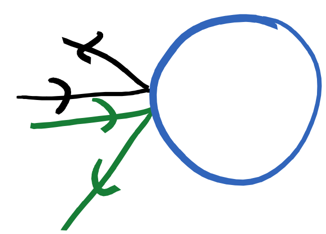
Feature Learning and BBI (in progress)

<u>To Do/in progress</u>: larger experiments including those requiring feature learning. ImageNet and variants in progress modulo resource requirements.

<u>Theory/intuition</u>: Chaos (with or without bounces) => diverging trajectories => feature learning even for `standard'/NTK initialization choices. cf Roberts/Yaida (criticality, large-width RG and minimal models), Yang/Hu (initialization enhancing hidden updates)

Compared to situation with hidden layers not updating (SGD at infinite width with NTK initialization), our chaotic dynamics contains diverging trajectories introducing $\Delta\theta_{hidden}$





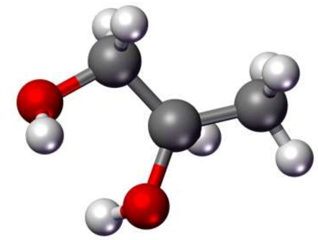
Computational Quantum Chemistry

(in progress w/Zhiyong Zhang, Stanford data science/nwchem developer)

Other analyses (sampling)/comparisons to additional global optimizers

(w/Uros Seljak)

Ongoing work:



- Quantum Chemistry (with Zhang)
  - Find the minimum energy configuration of a molecule

    F = binding energy < 0 $\implies$ requires ΔV
    - Automatic tuning tested successfully

- Larger scale Machine Learning experiments (with Kunin)
  - Exploit the volume formula from frictionless dynamics for better generalization

- Efficient sampling from a function exp(-F) (with Robnik, Seljak)
  - Reverse engineer g(V) such that

    Vol(phase space)= $\int d^n\Pi \int d^n\Theta \, \exp(-\mathrm{F})\delta\big(E - H(\Theta, \Pi)\big) \propto \int d^n\Theta \exp(-F)$

  - In contrast to Hamiltonian Monte Carlo, no momentum sampling needed

Future directions:

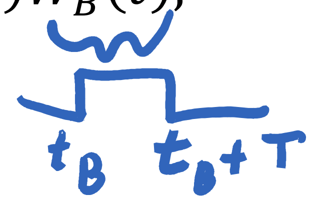
- Feature learning theory and experiment
  - Bounces along the directions of hidden layer parameters

Happy 60^{th}!

Extra Slides

Noisy case (mini-batches):

$$V(\theta(t),t) = \sum_B V^B(\{x\}_B,\theta)W_B(t), \qquad V_{full} = \sum_{\{x\}_B} V^B \quad e.g.\ V^B > 0\ \forall B$$



Time dependent potential (nonetheless we renormalize to the original E).
One can think of a given batch trajectory as deterministic.
Retains the main features:

- Cannot stop at local minimum (V>0)
- Will stop near V=0 due to speed limit

Also interesting to study ensemble averages, generalized Brownian motion:
Yaida ’18,… . Kunin Sagastuy-Brena, Gillespie, Tanaka, Ganguli, Yamins ‘21

BI: …+ $\mathrm{d}\frac{\langle\theta^2\rangle}{dt} \sim \langle\dot{\theta}^2\rangle < V$ (speed limit) vs Brownian motion: $\frac{\mathrm{d}\langle\theta^2\rangle}{dt} \propto \langle\dot{\theta}^2\rangle$

Application to PDEs in string compactifications

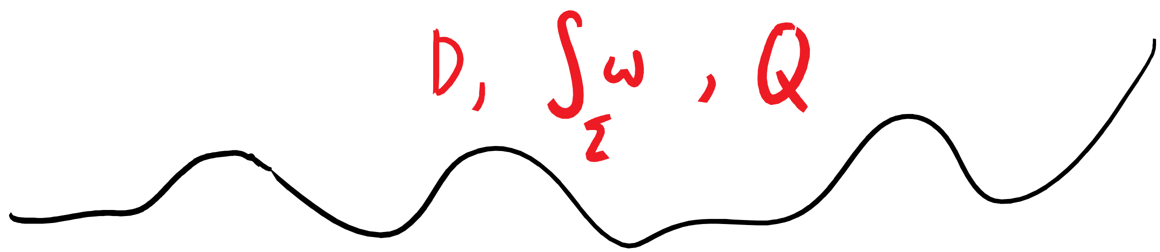
(w/G.B. De Luca, G. Torroba '21), cf e.g.

L.B. Anderson, M. Gerdes, J. Gray, S. Krippendorf, N. Raghuram and F. Ruehle, *Moduli-dependent Calabi-Yau and SU(3)-structure metrics from Machine Learning*, *JHEP* **05** (2021) 013 [2012.04656].

M.R. Douglas, S. Lakshminarasimhan and Y. Qi, *Numerical Calabi-Yau metrics from holomorphic networks*, 2012.04797.

V. Jejjala, D.K. Mayorga Pena and C. Mishra, *Neural Network Approximations for Calabi-Yau Metrics*, 2012.15821.

String theory & Cosmology

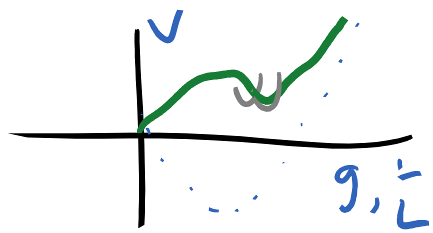


Structure of dS and inflation in string theory

--model-dependent UV sensitive observational tests

--microphysics of dS quantum gravity

--targets and methods for modern numerical methods and machine learning



4d effective potential

Douglas '09

Mostly positive:
$D - D_c, -R^{D-4}, (Q_1 + a\, Q_2)^2, \ldots$
Intermediate negative:
O-planes, quantum

$$V_{eff}[g^{(D-4)},\ldots] = \frac{\ell_D^{D-2}}{2G_N^2}\frac{\int d^{D-4}y\,\sqrt{g^{(D-4)}}\,e^{-2\Phi}u^2|_c\left(-R^{(D-4)} - \frac{1}{4}\ell_D^{D-2}T^{\mu}_{\mu} - 3\left(\frac{\nabla u}{u}\right)^2\Big|_c\right)}{\left(\int d^{D-4}y\sqrt{g^{(D-4)}}\,e^{-2\Phi}u|_c\right)^2}$$

Net curvature

$u(y) = e^{2A(y)}$

$R_{sec} < 0$ rigid
(cf Trodden et al, Saltman-ES, DLST)

$R_{ij} = 0$ CY
(cf KKLT, LVS…)

$$ds^2 = e^{2A(y)}ds^2_{dS_4} + e^{2B(y)}(g_{\mathbb{H}ij} + h_{ij})dy^i dy^j$$

u(y) satisfies GR constraint (its eq. of motion):

$$\left(-\nabla^2 - \frac{1}{3}\left(-R^{(D-4)} - \frac{1}{4}\ell_D^{D-2}T^{\mu}_{\mu}\right)\right)u = -\frac{C}{6}$$

Like a Schrodinger problem for
$C\ell^2 \sim H^2\ell^2 \ll 1$

$$\Rightarrow \quad V_{eff} = \frac{C}{4G_N} = \frac{R^{(4)}_{\rm symm}}{4G_N}\,.$$

Warp factor stabilizes runaway negativity (e.g. $-B'^2$)

dS examples stabilizing extra dimensions:

Reviews of various aspects: Polchinski, Baumann/McAllister, Douglas/Kachru, Denef, Frey, Hebecker; ES TASI '16, …

- Non-perturbative stabilization

  --GKP '01/KKLT '03 and many followups, e.g.
  --large volume scenario

  Sub-KK scale SUSY breaking

- Power-law stabilization

  --(D-Dc), O-planes, flux, asymmetric orbifold (large-D expansion) '01-'02
  (…other examples…)
  --hyperbolic space, Casimir, flux '21

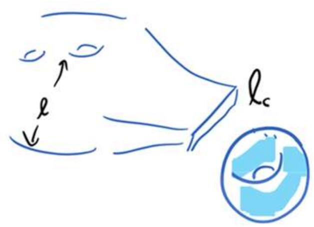
  -- RG logs & powers Burgess/Quevedo '22

  --including explicit uplifts of AdS/CFT
  [D1-D5 theory -> dS3 '10,
  M2 brane theory -> dS4 '21]

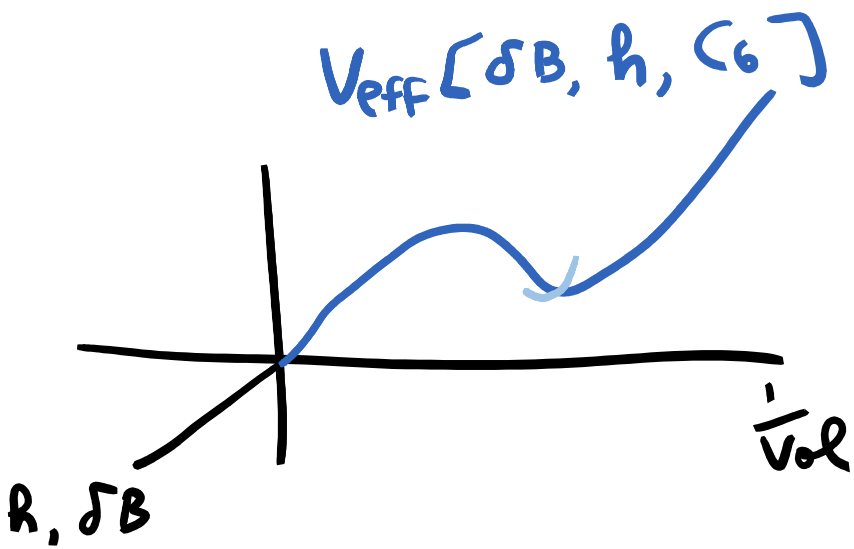
  ≥KK scale SUSY breaking

Weak-coupling EFT/large-N/Large-D/small W_0 control.
Ongoing studies of internal equations of motion in various cases & models, including ones with significant gradients e.g. Cordova et al, …

Curved internal dim's: recent mechanism for Λ from string/M theory



M theory (EFT: 11d SUGRA) on explicit infinite discrete family of finite-volume hyperbolic spaces with $\int -R - 3u'^2 \ll -\int R$ **parametrically**, automatically-generated Casimir energy, 7-form flux yields immediate volume stabilization and approximate piecewise solution dressed with warp & conformal variations, small residual tadpoles.



Strong positive Hessian contributions from **hyperbolic rigidity** and from **warping** (redshifting) effects on conformal factor and on Casimir energy.

Douglas '09

4d effective potential

$\ell_{11}^9 \rho_c(R_c) \sim -\frac{\ell_{11}^9}{R_c^{11}}$

$$V_{eff}[g^{(7)}, C_6] = \frac{\ell_{11}^9}{2G_N^2} \frac{\int d^7y \sqrt{g^{(7)}} u^2|_c \left(\overbrace{[-R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 \Big|_c]}^{\text{net curvature term}} - \frac{1}{4}\ell_{11}^9 T^{(Cas)\mu}{}_{\mu} + \frac{1}{2}|F_7|^2 \right)}{\left(\int d^7y \sqrt{g^{(7)}} u|_c\right)^2}$$

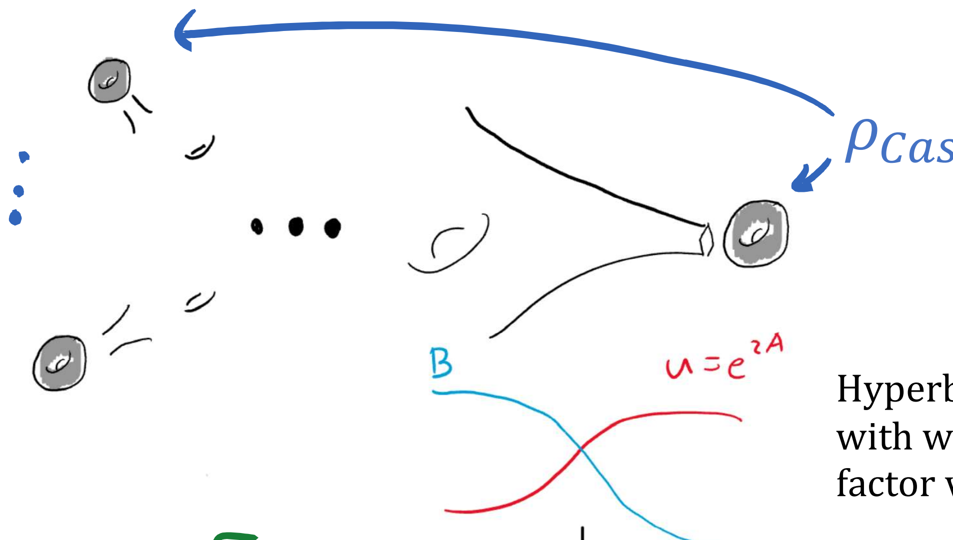
$$ds^2 = e^{2A(y)} ds^2_{dS_4} + e^{2B(y)} (g_{\mathbb{H}ij} + h_{ij}) dy^i dy^j \qquad\qquad u(y) = e^{2A(y)}$$

u(y) satisfies GR constraint (its equation of motion):

$$\left(-\nabla^2 - \frac{1}{3}\left(-R^{(7)} - \frac{1}{4}\ell_{11}^9 T^{(\mathrm{Cas})\mu}{}_{\mu} + \frac{1}{2}|F_7|^2\right)\right) u = -\frac{C}{6}$$

Like a Schrodinger problem for $C\ell^2 \sim H^2\ell^2 \ll 1$

$$\Rightarrow \quad V_{eff} = \frac{C}{4G_N} = \frac{R^{(4)}_{\mathrm{symm}}}{4G_N}\,.$$



Hyperbolic manifold dressed with warp and conformal factor variations

Tune small to compete with Casimir with $\ell_{11} \ll R_c \ll \ell$

$$\left[\, -R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 < 0 \;\middle|\; -R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 > 0 \,\right]$$

warp & conformal factor eoms $\Rightarrow$

$$-R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 = 4\ell_{11}^9|\rho_C| - \frac{C}{u} - \frac{5}{2}F_7^2$$

Douglas Kallosh '10

$$a = \frac{\int \sqrt{g^{(7)}} u^2|_c [-R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 \Big|_c]}{\int \sqrt{g^{(7)}} u^2|_c \, 42/\ell^2} \qquad << 1 \;:$$

$$-R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 = 4\ell_{11}^9 |\rho_C| - \frac{C}{u} - \frac{5}{2} F_7^2$$

- If a is too large, increase volume of non-Casimir regions (e.g. via short filled cusps or covers k-fold -> (k+1)-fold)

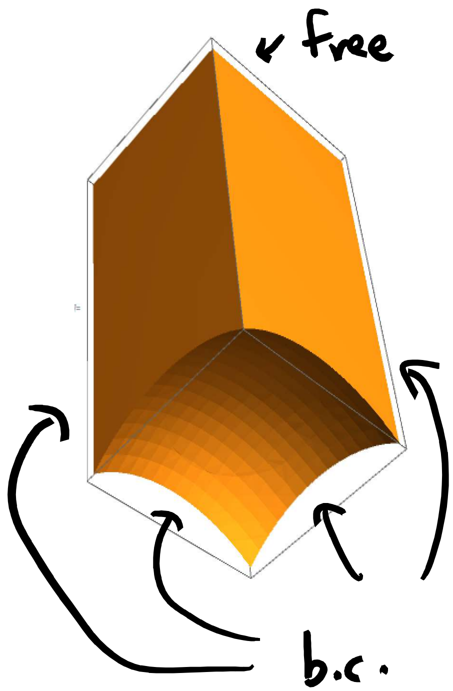
- If a is too small, reduce flux quantum number

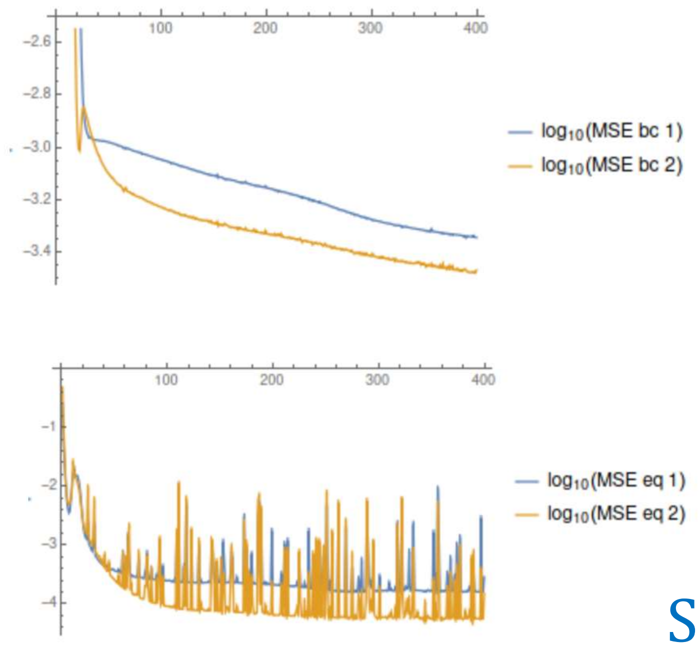
Work with simple concrete hyperbolic manifolds with comparable cusp and bulk volumes Italiano et al ‘20. Explicit radial solution illustrates $a << 1$.

Parametric suppression of residual tadpoles.

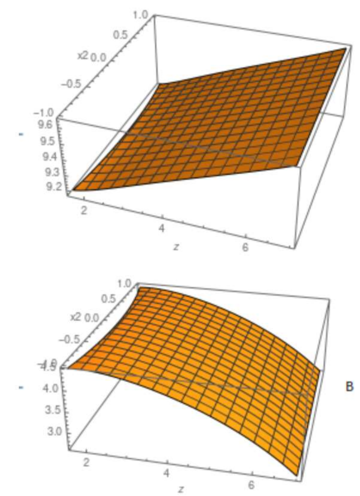
Numerical PDE solutions yields further details of solutions
(interesting for exploring beyond perturbative regime)

H_3 warmup example:





Loss



Slice of approximate solution for warp and conformal factors

Numerical study of this class of compactifications is fully specified and well-posed, including the stress-energy sources relevant for dS:

- H_7/Γ explicit projection of H_7, can also be constructed as gluing of explicit set of polygons.
- $\Gamma \Rightarrow$Casimir energy
- F_7 solution explicit in terms of metric
- Parametric limit(s) involving covers and filled cusps to compare to.

For ML, can consider PDE's, V_{eff}, or slow roll functionals ϵ_V, η_V as natural loss functions to explore.