Lecture Notes CAO 2024

4 Proximal gradient method

Proximal mapping

The prox-operator of a convex function h is defined as

$$\text{prox}_h(x) = \arg\min_u (h(u) + \frac{1}{2}\|u - x\|^2).$$

4.1 Examples of prox-operators

- $h(x) = 0$: $\text{prox}_h(x) = x$.
- Let C be a closed convex set. Let

$$\delta_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

  the *indicator function* of C. Then $\text{prox}_{\delta_C} = P_C$ the projection on C.
- $h(x) = \|x\|_1$: $\text{prox}_h = T_1$ where T_λ is the *soft-threshold* operator

$$T_\lambda(x)_i = \begin{cases} x_i - \lambda & x_i \geq \lambda \\ 0 & -\lambda \leq x_i \leq \lambda \\ x_i + \lambda & x_i \leq -\lambda. \end{cases}$$

4.2 Proximal gradient

Now we turn our attention to (unconstrained) optimization problems with more structure. Assume $f(x) = g(x) + h(x)$ where

- g is convex, differentiable, $\text{dom}\, g = \mathbb{R}^n$
- h convex with inexpensive prox-operator.

The Proximal Gradient method

To minimize f: choose an initial point x_0 and repeat

$$x_{k+1} = \text{prox}_{\eta_k h}(x_k - \eta_k \nabla g(x_k)),\ k = 0, 1, \ldots$$

where step size $\eta_k > 0$ constant or determined by line search.

Remark 1. *Proximal-gradient can start at infeasible x_0, but for all $k > 0$ we always have $x_k \in \text{dom}\, f = \text{dom}\, h$.*

4.2.1 Interpretation I.

$$x^+ = \text{prox}_{\eta h}(x - \eta \nabla g(x)).$$

From definition of prox-operator,

$$\begin{aligned} x^+ &= \arg\min_u (\eta h(u) + \frac{1}{2}\|u - x + \eta \nabla g(x)\|^2) \\ &= \arg\min_u (h(u) + \frac{1}{2\eta}(\|u - x\|^2 + 2\eta(u - x)^T \nabla g(x) + \eta^2 \|\nabla g(x)\|^2) \\ &= \arg\min_u (h(u) + \frac{1}{2\eta}\|u - x\|^2 + (u - x)^T \nabla g(x)) \\ &= \arg\min_u (h(u) + \frac{1}{2\eta}\|u - x\|^2 + (u - x)^T \nabla g(x) + g(x)) \end{aligned}$$

x^+ minimizes $h(u)$ plus a simple quadratic approximation of $g(u)$ around x.

4.2.2 Examples of proximal gradient

- **Gradient method:** Take $h(x) = 0$. Iterate

  $$x^+ = x - \eta \nabla g(x)$$

  to solve

  $$\min_x g(x).$$

- **Gradient projection method:** Take $h(x) = \delta_C(x)$. Iterate

  $$x^+ = P_C(x - \eta \nabla g(x))$$

  to solve

  $$\min_x \{g(x) + \delta_C(x)\} = \min_{x \in C} g(x).$$

- **Proximal method:** Take $g(x) = 0$. Iterate

  $$x^+ = \text{prox}_h(x)$$

  to solve

  $$\min_x h(x).$$

- **Soft-thresholding:** Take $h(x) = \lambda \|x\|_1$. Iterate

  $$x^+ = T_{\lambda\eta}(x - \eta \nabla g(x))$$

  to solve

  $$\min_x \{g(x) + \lambda \|x\|_1\}. \text{ (norm-1 penalized)}$$

4.3 The proximal mapping

The proximal mapping is a key element not only on the proximal gradient method, but also in the Douglas-Rachford and ADMM methods.

Proximal mapping

If h convex and closed (has closed epigraph), then

$$\text{prox}_h(x) = \arg\min_u \{h(u) + \tfrac{1}{2}\|u - x\|^2\}$$

exist and is unique for all x (i.e. prox_h is well defined).

Lemma 1. *Let h be convex and closed. Then $u = \text{prox}_h(x)$ if and only if $x - u \in \partial h(u)$.*

Proof. Follows from the optimality conditions. □

4.3.1 Properties of the proximal mapping

Lemma 2. *Let h be convex and closed.*

1. *Monotonicity of the subgradient:*

$$\text{Let } g_x \in \partial h(x) \text{ and } g_y \in \partial h(y). \text{ Then } (g_x - g_y)^\intercal (x - y) \geq 0.$$

2. prox_h *is* Firmly non - expansive*:*

$$(\text{prox}_h(x) - \text{prox}_h(y))^T (x - y) \geq \|\text{prox}_h(x) - \text{prox}_h(y)\|^2.$$

3. prox_h *is* non - expansive *(1 - Lipschitz continouos)*

$$\|\text{prox}_h(x) - \text{prox}_h(y)\| \leq \|x - y\|.$$

Proof. Exercise □

4.3.2 Interpretation of proximal gradient II

Now we give an interpretation of the proximal gradient based on fix-point operators.

To solve the problem $\min\{h(x) + g(x)\}$ is equivalent to find x such that $0 \in \partial h(x) + \nabla g(x)$, or equivalently to find x such that $-\nabla g(x) \in \partial h(x)$. This last equation can be written as $(x - \eta\nabla g(x)) - x \in \partial(\eta h)(x)$. And by Lemma 1 this is equivalent to $x = \text{prox}_{\eta h}(x - \eta\nabla g(x))$. Lets define $F_{PG}(x) := \text{prox}_{\eta h}(x - \eta\nabla g(x))$. We need to solve $F_{PG}(x) = x$, that is to find a fixed-point of F_{PG}.

There are several methods to find fixed-points. The most common being fix-point iteration. I.e. take a initial x_0 and iterate $x_{k+1} = F(x_k)$. Notice that if the sequence converges, the limit point is a fix-point of F. There are various properties of F that guarantee the convergence of the fix-point iteration. The most common being the contracting property.

Contracting property

$F : \mathbb{R}^n \to \mathbb{R}^n$ is contracting if there is $\rho < 1$ such that $\|F(x) - F(y)\| \leq \rho\|x - y\|$ for all x and y.

Theorem 1 (Banach Fixed Point Theorem)**.** *Let F be a contraction mapping. Then F admits a unique fixed-point and the fix-point iteration algorithm always converges to such fixed-point.*

Notice that we can define a similar operator for the gradient method $F_G(x) := x - \varepsilon\nabla g(x)$, and similarly we will define operator for Douglas-Rachford and for ADMM. These operators are in general not contracting. But they are firmly non-expansive. In the next section we will review some tools from fixed-point theory that will allow us to show that the fix-point method convergence under this weaker assumption. Also we will develop tools to analyse convergence rates.

Bibliography

- A. Beck, First-Order Methods in Optimization (2017), sect. 10.4 – 10.6.
- A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences (2009).
- A. Beck and M. Teboulle, Gradient-based algorithms with applications to signal recovery, in: Y. Eldar and D. Palomar (Eds.), Convex Optimization in Signal Processing and Communications (2009).
- Yu. Nesterov, Lectures on Convex Optimization (2018), sect. 2.2.32.2.4.
- B. T. Polyak, Introduction to Optimization (1987), sect. 7.2.1.

5 Fix-point operators

Many iterative algorithms in optimization can be interpreted as fix-point iteration, which is used to solve fix point equations. Namely, let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous operator and let $x_0 \in \mathbb{R}^n$ be given. Define iteratively the sequence $x_{k+1} = F(x_k)$ for $k \geq 0$, and assume x_k converges to $\hat{x}$. We have then,

$$F(\hat{x}) = F(\lim_{k\to\infty} x_k) = \lim_{k\to\infty} F(x_k) = \lim_{k\to\infty} x_{k+1} = \hat{x},$$

i.e. $\hat{x}$ is a fixed-point of F. We would use $\mathcal{F}_F := \{x \in \mathbb{R}^n : x = F(x)\}$ to denote the set of fixed points of the operator F.

As mentioned in the previous section, some of the algorithms studied in previous sections fall into this idea.

Optimization methods as Fix-Point operators

The following are examples of optimization-related fix point operators

- $F_G(x) := x - \varepsilon\nabla g(x)$ (gradient descend),
- $F_G^C(x) := P_C(x - \varepsilon\nabla g(x))$ (projected gradient method),
- $F_{Prox}(x) := \text{prox}_{\nabla h}(x)$ (proximal method), and
- $F_{PG}(x) := \text{prox}_{\nabla h}(x - \varepsilon\nabla g(x))$ (proximal gradient method).

Lemma 3. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a given differentiable convex function, let $C \subset \mathbb{R}^n$ be a closed convex set and let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We have then*

1. $\mathcal{F}_{F_G} = \{x \in \mathbb{R}^n : \nabla g(x) = 0\} = \arg\min g(x)$,

2. $\mathcal{F}_{F_G^C} = \{x \in C : g(x)^T(y-x) \geq 0 \text{ for all } y \in C\} = \arg\min_{x \in C} g(x)$, *and*
3. $\mathcal{F}_{F_{Prox}} = \{x \in \mathbb{R}^n : 0 \in \partial h(x)\} = \arg\min h(x)$.
4. $\mathcal{F}_{F_{PG}} = \{x \in \mathbb{R}^n : 0 \in \nabla g(x) + \partial h(x)\} = \arg\min g(x) + h(x)$.

Proof. Exercise. □

Remark 2. *Notice that the subgradient method is not (in a straightforward sense) a fix-point iteration algorithm.*

We will see later other operators which also have as fixed-points the solution to given optimization problems. Now, we turn our attention to the question of the convergence of fix-point iteration.

5.1 (Linear) convergence of fix point operators

First, we give conditions for the convergence of fix-point operators (that is the convergence of fix point iteration for this operators). It is natural to look at the *expansiveness* of the operator. Namely to obtain convergence nearby points should be send nearby. Banach's theorem (Theorem 1) ensures the convergence of contracting operators. We concentrate then in the class of non-expansive operators.

Non-expansion

$F : \mathbb{R}^n \to \mathbb{R}^n$ is

- non-expansive if $\|F(x) - F(y)\| \leq \|x - y\|$ for all x and y,
- firmly non-expansive if $(F(x) - F(y))^T(x - y) \geq \|F(x) - F(y)\|^2$.
- α - averaged if $F(x) = (1-\alpha)x + \alpha N(x)$ where N is non-expansive,

Theorem 2 (characterization of averaged and non-expansive operators)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$.*

1. *F is firmly non-expansive if and only if it is $1/2$-averaged.*
2. *Let $\alpha \in (0,1)$. If F is α-averaged then for all $x, y \in \mathbb{R}^n$ we have*
$$\|F(x) - F(y)\|^2 + \tfrac{1-\alpha}{\alpha}\|(x - F(x)) - (y - F(y))\|^2 \leq \|x - y\|^2.$$
3. *Averaged operators are closed under composition.*

Non-expansive operators do not need to be convergent. Consider for instance a rotation. On the other hand, we have that averaged operators converge (see Thorem 3. That is, damped iteration of a non-expansive operator will converge (to one of its fixed points).

Lemma 4. *Let $\alpha \in (0,1)$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be α-averaged. Let $\hat{x} \in \mathcal{F}_F$ be a fixed-point of F. Then*
$$\tfrac{1-\alpha}{\alpha}\|F(x) - x\|^2 + \|F(x) - \hat{x}\|^2 \leq \|x - \hat{x}\|^2.$$

Proof. Follows from Theorem 2.2 by taking $y = \hat{x} = F(y)$. □

Theorem 3 (Convergence of averaged fix-point operators)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be an averaged operator, such that F has fixed points. Given $x_0 \in \mathbb{R}^n$, let $x_{k+1} = F(x_k)$ for $k = 0, 1, \dots$. Then, $x^* = \lim_{k\to\infty} x_k$ exists and $F(x^*) = x^*$.*

Proof. Take any $\hat{x} \in \mathcal{F}_F$. By Lemma 4 for all k we have

$$\tfrac{1-\alpha}{\alpha}\|x_k - x_{k-1}\|^2 + \|x_k - \hat{x}\|^2 \leq \|x_{k-1} - \hat{x}\|^2, \tag{1}$$

and thus $\|x_k - \hat{x}\| \leq \|x_0 - \hat{x}\|$. Therefore, there exists $\bar{x} \in \mathbb{R}^n$ such that $x_{k_j} \to \bar{x}$ for some subsequence $\{x_{k_j} : j = 0, 1, \dots\}$ of $\{x_k : k = 0, 1, \dots\}$. We next show that indeed $\bar{x} \in \mathcal{F}_F$ and $x_k \to \bar{x}$.

Since $x_{k_j} \to \bar{x}$, the continuity of F implies that $x_{k_j+1} = F(x_{k_j}) \to F(\bar{x})$. In addition, by (1) we have $\frac{1-\alpha}{\alpha}\sum_{j=1}^{k}\|x_j - x_{j-1}\|^2 \leq \|x_0 - \hat{x}\|^2 - \|x_k - \hat{x}\|^2 \leq \|x_0 - \hat{x}\|^2$. Thus, $\|x_k - x_{k-1}\| \to 0$. Therefore, $x_{k_j} \to \bar{x}$ implies $x_{k_j+1} \to \bar{x}$, that is $\bar{x} = F(\bar{x})$. Since our argument holds for any $\hat{x} \in \mathcal{F}_F$, in particular it holds for $\hat{x} = \bar{x}$ and so $\|x_k - \bar{x}\|$ is monotonically decreasing. Since $\|x_{k_j} - \bar{x}\| \to 0$ it follows that $\|x_k - \bar{x}\| \to 0$, that is, $x_k \to \bar{x}$. □

It can also be shown that under the same assumptions of Theorem 3 the rate of convergence of $\|x_{k+1} - x_k\| \to 0$ is $O(1/\sqrt{k})$ [**?**] (see also [Theorem 1 in **?**]). As several well-known optimization algorithms are equivalent to the FPI of averaged operators, by Theorem 3, we obtain sublinear convergence of all these algorithms. Next we tackle the question of linear convergence. In Theorem 4 linear convergence for the fix-point iterates is shown when the following *error bound condition* holds for a suitable $R > 0$:

There exists $K_F > 0$ such that

$$\operatorname{dist}(x, \{x : F(x) = x\}) \leq R \text{ implies } \operatorname{dist}(x, \{x : F(x) = x\}) \leq K_F \cdot \|F(x) - x\|. \tag{2}$$

Theorem 4 (Error bound implies Linear convergence[van Treek, Pena, Vera, Zuluaga 2024)**.** *] Let $\alpha \in (0, 1)$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous α-averaged operator with fixed-points. Assume the error bound condition (2) holds for $R > 0$. Let x_0 be given and let $x_{k+1} = F(x_k)$ be for $k = 0, 1, \dots$. Then $x_k \to \bar{x} \in \mathcal{F}_F$ linearly. In particular,*

$$\operatorname{dist}(x_{k+1}, \mathcal{F}_F) \leq \left(1 - \tfrac{1-\alpha}{2\alpha K_F^2}\right)\operatorname{dist}(x_k, \mathcal{F}_F) \text{ and } \|x_{k+1} - \bar{x}\| \leq \left(1 - \tfrac{(1-\alpha)^2}{8\alpha^2 K_F^4}\right)\|x_k - \bar{x}\|,$$

Proof. Proof. By Theorem 3, fixed-point iteration of averaged operators converge. Thus, we know the sequence x_k converges to a point $\bar{x} \in \mathcal{F}_F$. Let $R > 0$ be such that the error bound condition (2) holds. Let k_0 be the smallest $k \geq 0$ such that $\operatorname{dist}(x_k, \mathcal{F}_F) \leq R$. Such k_0 exists as x_k is converging to a point in $\mathcal{F}_F$. Let

$$\rho = \left(1 - \tfrac{1-\alpha}{\alpha K_F^2}\right)^{\frac{1}{2}}, \tag{3}$$

we show that for all $k \geq k_0$,

$$\operatorname{dist}(x_{k+1}, \mathcal{F}_F) \leq \rho \operatorname{dist}(x_k, \mathcal{F}_F). \tag{4}$$

First, notice that for all $k \geq k_0$ we have $\operatorname{dist}(x_k, \mathcal{F}_F) \leq R$ as averaged implies non-expansive. Now, for each $k \geq k_0$, let $\bar{x}_k$ be a solution to $\inf_{x\in\mathcal{F}_F}\|x_k - x\|$. As $\mathcal{F}_F$ is closed, we have $\bar{x}_k \in \mathcal{F}_F$, and $\|x_k - \bar{x}_k\| = \operatorname{dist}(x_k, \mathcal{F}_F)$.

Fix $k \geq k_0$. By Lemma 4,

$$\|x_{k+1} - \bar{x}_k\|^2 + \tfrac{1-\alpha}{\alpha}\|x_k - x_{k+1}\|^2 \leq \|x_k - \bar{x}_k\|^2. \tag{5}$$

By the error bound condition (2),

$$\|x_k - \bar{x}_k\| = \operatorname{dist}(x_k, \mathcal{F}_F) \leq K_F \|x_k - x_{k+1}\|. \tag{6}$$

Combining (5) with (6) gives

$$\|x_{k+1} - \bar{x}_{k+1}\|^2 \leq \|x_{k+1} - \bar{x}_k\|^2 \leq \|x_k - \bar{x}_k\|^2 - \tfrac{1-\alpha}{\alpha}\|x_k - x_{k+1}\|^2 \leq \left(1 - \tfrac{1-\alpha}{\alpha K_F^2}\right)\|x_k - \bar{x}_k\|^2.$$

By taking square roots in both sides we obtain (4). By (3) we obtain $\rho \leq 1 - \frac{1-\alpha}{2\alpha K_F^2}$ and the first part of the statement follows.

If $\mathcal{F}_F$ is a singleton, then $\bar{x}_k = \bar{x}$ for all k, and (4) implies that $\|x_{k+1} - \bar{x}\| \leq \rho \|x_k - \bar{x}\|$. So, $x_k \to \bar{x}$ linearly with rate ρ. Otherwise, note that by (4), for any $k, j = 1, 2, \ldots$

$$\|\bar{x}_{k+j+1} - \bar{x}_{k+j}\| \leq \|\bar{x}_{k+j+1} - x_{k+j}\| + \|x_{k+j} - \bar{x}_{k+j}\| \leq 2\|x_{k+j} - \bar{x}_{k+j}\| \leq 2\rho^{j+1}\|x_k - \bar{x}_k\|.$$

In addition, $\bar{x}_k \to \bar{x}$, since $x_k \to \bar{x}$ and $\|x_k - \bar{x}_k\| \to 0$. Thus,

$$\begin{aligned}\|x_{k+1} - \bar{x}\| &\leq \|x_{k+1} - \bar{x}_{k+1}\| + \sum_{j=1}^{\infty} \|\bar{x}_{k+j+1} - \bar{x}_{k+j}\| \\ &\leq \left(\rho + 2\sum_{j=1}^{\infty} \rho^{j+1}\right)\|x_k - \bar{x}_k\| \leq \tfrac{\rho(1+\rho)}{1-\rho}\|x_k - \bar{x}_k\|. \end{aligned} \tag{7}$$

Therefore, combining (5), (6) and (7), we find that

$$\begin{aligned}\|x_{k+1} - \bar{x}\|^2 &\leq \|x_k - \bar{x}\|^2 - \tfrac{1-\alpha}{\alpha}\|x_k - x_{k+1}\|^2 \leq \|x_k - \bar{x}\|^2 - \tfrac{1-\alpha}{\alpha K_F^2}\|x_k - \bar{x}_k\|^2 \\ &\leq \|x_k - \bar{x}\|^2 - \tfrac{(1-\alpha)(1-\rho)}{\alpha\rho(1+\rho)K_F^2}\|x_{k+1} - \bar{x}\|^2.\end{aligned}$$

Equivalently,

$$\|x_{k+1} - \bar{x}\| \leq \left(1 + \tfrac{(1-\alpha)(1-\rho)}{\alpha(1+\rho)\rho K_F^2}\right)^{-\frac{1}{2}}\|x_k - \bar{x}\|.$$

To finish the proof we simplify the above expression. Let $u = \frac{1-\alpha}{\alpha K_F^2}$. By (3) we have $u = 1 - \rho^2$ and thus $\left(1 + \frac{(1-\alpha)(1-\rho)}{\alpha(1+\rho)\rho K_F^2}\right)^{-\frac{1}{2}} = \left(1 + \frac{(1-\rho)^2}{\rho}\right)^{-\frac{1}{2}}$.

Using $\frac{(1-\rho)^2}{\rho} \geq (1-\rho)^2 \geq \frac{u^2}{4}$ and $(1+\theta)^{-\frac{1}{2}} \leq 1 - \frac{1}{4}\theta$, for any $\theta \in (0, 1)$ we obtain

$$\left(1 + \tfrac{(1-\alpha)(1-\rho)}{2\alpha\rho K_F^2}\right)^{-\frac{1}{2}} \leq 1 - \tfrac{(1-\alpha)^2}{8\alpha^2 K_F^4}.$$

□

Theorem 4 allows us to recast the problem of proving and estimating the rate of linear convergence as the problem of computing the constant K_F in the generic error bound condition (2). We will illustrate this in Section 6 by estimating the rate of linear convergence of optimization algorithms.

By looking closely to the proof of Theorem 4 we can see that we have proven a stronger statement than the linear convergence of the FPI. We prove that if the error bound condition (2) holds for $R > 0$ then the distance to the set of fixed points decreases linearly when F is applied to any point in $\{x : \operatorname{dist}(x, \mathcal{F}_F) \leq R\}$. Next, in Lemma 1 we prove the converse of this statement, showing that the error bound condition is necessary to obtain this form of linear convergence.

Proposition 1 (Linear convergence implies error bound [van Treek, Pena, Vera, Zuluaga 2024]**.** *] Let $F : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathcal{F}_F \neq \emptyset$. Let $R > 0$ and $\rho_F < 1$ be such that*

$$\operatorname{dist}(x, \mathcal{F}_F) \leq R \text{ implies } \operatorname{dist}(F(x), \mathcal{F}_F) \leq \rho_F \operatorname{dist}(x, \mathcal{F}_F). \tag{8}$$

Then, the error bound condition (2) *holds for R, with constant $K_F = \frac{1}{1-\rho_F}$*

Proof. Given $x \in S$, let $\bar{x}, \hat{x} \in \mathcal{F}_F$ such that $\operatorname{dist}(F(x), \mathcal{F}_F) = \|F(x) - \hat{x}\|$ and $\operatorname{dist}(x, \mathcal{F}_F) = \|x - \bar{x}\|$. By (8),

$$\|F(x) - \hat{x}\| \leq \rho_F \|x - \bar{x}\| \leq \rho_F \|x - \hat{x}\| \leq \rho_F (\|x - F(x)\| + \|F(x) - \hat{x}\|).$$

Reorganizing the terms we obtain the statement. □

Theorem 4 and Proposition 1 relate the rate of convergence ρ_F of the FPI to the error bound constant K_F.

Corollary 1 (Error-bound condition and linear convergence bounds [van Treek, Pena, Vera, Zuluaga 2024]**.** *] Let $\alpha \in (0, 1)$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous α-averaged operator with $\mathcal{F}_F \neq \emptyset$. Let $R > 0$ and define $\tilde{\rho}_F \in [0, 1]$ and $\tilde{K}_F \in \mathbb{R} \cup \{\infty\}$ as*

$$\tilde{\rho}_F = \sup_x \frac{\operatorname{dist}(F(x), \mathcal{F}_F)}{\operatorname{dist}(x, \mathcal{F}_F)} \text{ and } \tilde{K}_F = \sup_x \frac{\operatorname{dist}(x, \{x : F(x) = x\})}{\|F(x) - x\|},$$

the tightest constants such that linear convergence result (8)*, respectively the error-bound condition* (2) *hold. We have the following equivalent relations,*

$$1 - \frac{1}{\tilde{K}_F} \leq \tilde{\rho}_F \leq 1 - \frac{1 - \alpha}{2\alpha \tilde{K}_F^2},$$

and

$$\sqrt{\frac{1 - \alpha}{2\alpha(1 - \tilde{\rho}_F)}} \leq \tilde{K}_F \leq \frac{1}{1 - \tilde{\rho}_F}.$$

Bibliography

- *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* by Heinz H. Bauschke, Patrick L. Combettes. CMS Books in Mathematics (2011). Ch4 and ch27.
- Baillon, J. B. and Bruck, R. E. The rate of asymptotic regularity is $o(1/\sqrt{(n)})$. *Lecture Notes in Pure and Applied Mathematics*, 178:51–81 (1996).
- Cominetti, R., Soto, J. A., and Vaisman, J. On the rate of convergence of krasnosel'skiĭ-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*, 199(2):757–772 (2014).
- N. Parikh and S. Boyd. *Proximal Algorithms.*Foundations and Trends in Optimization Vol. 1, No. 3 (2013) 123–231.
- K. van Treek, J.F. Pena, J.C. Vera and L.F. Zuluaga. *Equivalence between Linear Convergence and Error Bounds for Optimization Algorithms.* Working paper. Available by request.

6 Convergence of optimization methods

Here we apply the results of previous sections to show the convergence of different optimization methods. Notice that the gradient method, the projected gradient method and the proximal method are all particular cases of the proximal gradient. thus we only need to analyse the proximal gradient.

An important class of functions for us is the family of continuous *piecewise linear-quadratic* (PLQ) functions. A closed convex function is PLQ if there exists a finite polyhedral partition of its domain such that in each of the parts the function is quadratic. Notice that indicator of polyhedral sets, quadratic functions and piece-wise linear functions are PLQ. Also, the class of PLQ functions is closed under sums and under composition with piecewise linear functions. Optimization of PLQ functions encompasses several important optimization models such as linear optimization, (convex) quadratic optimization, least squares, and some of its variations such as LASSO, elastic net, and support vector machines (SVM).

Lemma 5. *Let f be a closed convex continuous PLQ function. Then*

1. prox_f *is piece-wise linear.*
2. *If f is differentiable, ∇f is piece-wise linear.*

Proof. Exercise. □

6.1 Convergence of the proximal gradient method

We concentrate now on the proximal gradient method, by analysing the operator $F_{PG}(x) := \operatorname{prox}_{\eta h}(x - \eta \nabla g(x))$.

Assumption 1. *In this section we assume $f(x) = g(x) + h(x)$ where*

- *optimal value $f^* = \inf f(x)$ is finite and attained at x^**
- *h is closed and convex*
- *g is convex, differentiable,* $\operatorname{dom} g = \mathbb{R}^n$
- *∇g is L-Lipschitz continuous, with $L > 0$:*

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\| \text{ for all } x, y \in \operatorname{dom} g$$

- $0 < \eta \leq 1/L$.

Lemma 6. *Under Assumption 1,*

1. $(\nabla g(x) - \nabla g(y))^T (x - y) \geq \frac{1}{L}\|\nabla g(x) - \nabla g(y)\|^2$ *for all* $x, y \in \operatorname{dom} g$
2. $I - \eta \nabla g$ *is* $1/2$*-averaged*
3. prox_h *is* $1/2$*-averaged.*

Proof. part 1 is equivalent to the L-Lipschitz continuity. Parts 2 is left as exercise. Part 3 follows by combining Lemma 2 and Theorem 2. □

Proposition 2. *Under Assumption 1, The proximal gradient operator F_{PG} is $3/4$-averaged.*

Proof. By Lemma 6 is enough to show that the composition of two 1/2-averaged operators is 3/4-averaged. This is left as an exercise. □

Corollary 2. *Under Assumption 1, the proximal gradient method converges to a minimizer of f. Moreover, if g and h are PLQ then the convergence is R-linear.*

Proof. From Proposition 2 F_{PG} is 3/4-averaged. The statement follows by Theorems 3 and 4. □

6.2 Douglas Rachford

As second example we consider the Douglas-Rachford algorithm. Given h and g closed convex functions, let $F_{DR} : \mathbb{R}^n \to R^n$ be defined by $F_{DR}(w) = w + \text{prox}_h(2\,\text{prox}_g(w) - w) - \text{prox}_g(w)$ for all $w \in R^n$.

Assumption 2. *In this section we assume $f(x) = g(x) + h(x)$ where*

- *optimal value $f^* = \inf f(x)$ is finite and attained at x^**
- *h and g closed and convex.*

First, we do some rewriting. By Lemma 2 the prox operator is firmly non-expansive. By Theorem 2 we obtain that R_h defined by $R_h(x) = 2\,\text{prox}_h(x) - x$ is non-expansive. Notice that we can write then $F_{DR}(w) = \frac{1}{2}(w + R_h(R_g(w)))$, which is a 1/2-averaged operator.

Corollary 3. *Assume 2. Let $w_0 \in R^n$. Let $w_{k+1} = F_{DR}(w_k)$ for all $k \geq 0$. Then $w_k \to \bar{w} \in \mathcal{F}_{F_{DR}}$. And $g(w_k) \to \bar{x} = g(\bar{w})$, where $\bar{x} \in \arg\min f(x)$. Moreover, if g and h are PLQ then the convergence is linear.*

Proof. Follows by Theorems 3 and 4. The details are left as an exercise. □

Bibliography

- *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* by Heinz H. Bauschke, Patrick L. Combettes. CMS Books in Mathematics (2011). Ch4 and ch27.
- N. Parikh and S. Boyd. *Proximal Algorithms.*Foundations and Trends in Optimization Vol. 1, No. 3 (2013) 123–231.
- K. van Treek, J.F. Pena, J.C. Vera and L.F. Zuluaga. *Equivalence between Linear Convergence and Error Bounds for Optimization Algorithms.* Working paper. Available by request.