Nonsmooth Convex Optimization Methods – Part II

Franck Iutzeler

November 12, 2021

Contents

CHAPTER 1 CONVEXITY

CONVEXITY is at the heart of optimization. This is notably due to the unicity of projections onto convex sets and the direct link between critical points and minimums for convex functions.

In this chapter, we will first study convex *sets*, then convex *functions*.

1.1 CONVEX SETS

1.1.1 Motivation: Projecting onto a closed set

Similarly to orthogonal projections onto affine subspaces, we can define projection on *nonempty closed* sets.[1]

Thus, let us consider a non-empty closed set C and investigate the problem

$$\inf_{x \in C} F_y(x) \coloneqq \frac{1}{2}\|y - x\|^2 \tag{1.1}$$

which intuitively amounts to projecting y onto C.

First, take $u \in C$, and define $S \coloneqq \{x \in \mathbb{R}^n : \|y - x\|^2 \leq \|y - u\|^2\}$. Then, the problem (1.1) is equivalent to

$$\inf_{x \in C \cap S} F_y(x) \coloneqq \frac{1}{2}\|y - x\|^2 \tag{1.2}$$

where $C \cap S$ is a closed compact set. Projecting thus amounts to minimizing a continuous function over a closed compact set, which always admits a solution, as per the following lemma.

Lemma 1.1. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semi-continuous function (or in particular, a continuous function) and let S be a closed compact set. Then, there is some $x^\star \in S$ such that $F(x^\star) = \inf_{x \in S} F(x)$.*

Proof. ([$\star$]) Since F is proper, it nevers takes the value $-\infty$ thus $\bar{\beta} \coloneqq \inf_{x \in S} F(x) > -\infty$. For a decreasing sequence of reals (β_n) with $\beta_n \to \bar{\beta}$, let us define the sequence of the $S_{\beta_n} = \{x : F(x) \leq \beta_n\}$. For any n, S_{β_n} is nonempty, closed, and included in $S_{\beta_{n-1}}$. Thus, the limit $S_{\bar{\beta}} = \{x : F(x) = \inf_{u \in S} F(u)\}$ is also nonempty and closed which gives the result. □

This grants the existence of a minimizer of (1.2), and thus of (1.1), ie. a projection on C. In particular, the inf above are actually min. However, the projection may not be unique, that is where convexity comes into play.[2]

[1] Nonempty: otherwise there is nothing to project onto. Closed: otherwise "the" closest point in a set from another point is not well-defined.

[2] The above enables us to show the existence of projections onto nonempty closed sets, but the projection may not be unique.

1.1.2 Convexity for sets

Let us now introduce the definition of a convex set.

Definition 1.2. A subset C of $\mathbb{R}^n$ is convex if and only if for any $x, u \in C$, $(1-\alpha)x+\alpha u \in C$ for any $\alpha \in (0,1)$.

The crucial property here is that any (weighted) average of points of a convex set belongs stay in the set. Equivalently, the set C is convex if and only if for any $(x_1, .., x_N) \in C^N$,

$$\sum_{i=1}^{N} \alpha_i x_i \in C \text{ for any } (\alpha_1, .., \alpha_N) \in \mathbb{R}_+^N \text{ with } \sum_{i=1}^{N} \alpha_i = 1,$$

where $\sum_{i=1}^N \alpha_i x_i$ is called a *convex combination* of $(x_1, .., x_N)$.

Examples of convex sets:

- Affine spaces $\{x : \langle s, x\rangle = r\}$
- Balls $\{x : \|x - s\| \leq r\}$
- Half spaces $\{x : \langle s, x\rangle \leq r\}$ and open half spaces $\{x : \langle s, x\rangle < r\}$
- Simplices $\{x : \sum_{i=1}^n x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i = 1, .., n\}$
- Intersections of convex sets $\cap_{i=1}^N C_i$

Examples of non-convex sets:

- Discrete sets (eg. $\{0\} \cup \{1\}$) or disjoint sets
- Spheres $\{x : \|x - s\| = r\}$
- Sets with "holes"

1.1.3 Projection on convex sets

Getting back to the projection problem (1.1)

$$\min_{x \in C} F_y(x) \coloneqq \frac{1}{2}\|y - x\|^2 \tag{1.3}$$

where $S \coloneqq \{x \in \mathbb{R}^n : \|y - x\|^2 \leq \|y - u\|^2\}$. Now, let us assume that C is additionally convex.

Suppose that $x_1^\star \neq x_2^\star$ are two distinct solutions of (1.3). Define $x_0^\star = (x_1^\star + x_2^\star)/2$, then

$$\begin{aligned} F_y(x_0^\star) &= \frac{1}{2}\|y - x_0^\star\|^2 = \frac{1}{2}\|(y - x_1^\star)/2 + (y - x_2^\star)/2\|^2 \\ &= \frac{1}{4}\|y - x_1^\star\|^2 + \frac{1}{4}\|y - x_2^\star\|^2 - \frac{1}{8}\|x_1^\star - x_2^\star\|^2 \\ &= \frac{1}{2}(F_y(x_1^\star) + F_y(x_2^\star)) - \frac{1}{8}\|x_1^\star - x_2^\star\|^2 \end{aligned}$$

thus $F_y(x_0^\star) < F_y(x_1^\star) = F_y(x_2^\star)$ which contradicts $x_1^\star \neq x_2^\star$ being two distinct solutions. Hence, the projection on a convex set is unique. We have shown the following lemma.

Lemma 1.3. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, there is a unique projection* $\text{proj}_C(y)$*, solution of* (1.3).

In fact, this unique projection can be characterized more precisely.

Theorem 1.4. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, $\mathrm{proj}_C(y)$ is the projection of y onto C if and only if*

$$\langle y - \mathrm{proj}_C(y), z - \mathrm{proj}_C(y)\rangle \leq 0 \text{ for all } z \in C.$$

Proof. Left as an exercise. See (Hiriart-Urruty and Lemaréchal, 1993b, Th. 3.1.1). □

1.2 Convex functions

The notion of convexity is as important for functions as for sets. Notably, this is the notion that will enable us to go from the (sub)gradient inequalities and local minimizers above to *global* minimizers.

1.2.1 Definition

A function is convex if and only if its *epigraph*[3] is convex. However, the following definition is much more direct.

[3] This is the set $\mathrm{epi}F \coloneqq \{(x,t) : F(x) \leq t\}$

Definition 1.5. A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if for any $x, u \in \mathrm{dom}\, F$, $F((1-\alpha)x + \alpha u) \leq (1-\alpha)F(x) + \alpha F(u)$ for any $\alpha \in (0,1)$.

More generally convex functions verify *Jensen's inequality*. For any convex combination $\sum_{i=1}^N \alpha_i x_i$,

$$F\left(\sum_{i=1}^N \alpha_i x_i\right) \leq \sum_{i=1}^N \alpha_i F(x_i).$$

Checking the definition directly may be possible but it is often simpler to rely on convexity-preserving operations:

- all norms are convex;
- a sum of convex functions is convex;
- affine substitution of the argument (if F is convex, $x \mapsto F(Ax+b)$ is convex for any affine map $Ax+b$);
- the (pointwise) maximum of convex functions is convex.

The most striking point of convex functions is that local minimizers are actually global.

Theorem 1.6. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. Then, every local minimizer of F is a (global) minimizer.*

Proof. Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function and let x be a local minimizer of F. Then, there is a ball of radius $\rho > 0$ such that $F(x) \leq F(u)$ for all $u \in \mathcal{B}(x,\rho)$. Take $y \in \mathbb{R}^n \setminus \mathcal{B}(x,\rho)$ and define $\alpha = \rho/\|y-x\|$. Since $0 < \rho < \|y-x\|$, we have $\alpha \in (0,1)$.

Now, let $z = (1-\alpha)x + \alpha y$, we have $\|z - x\| = \alpha\|x-y\| = \rho$ so $z \in \mathcal{B}(x,\rho)$.

Since $F(x) \leq F(u)$ for all $u \in \mathcal{B}(x,\rho)$, we have $F(x) \leq F(z) = F((1-\alpha)x + \alpha y) \leq (1-\alpha)F(x) + \alpha F(y)$ by convexity of F. Thus implies that $F(x) \leq F(y)$, thus x is a minimizer for F in $\mathcal{B}(x,\rho)$ and outside of it, thus a global minimizer. □

1.2.2 Proper lower-semicontinuous functions

Before studying differentiability, we will need to define the notions of domain, optimality, properness, and lower-semicontinuity.

For a function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$, we define its *domain* as $\operatorname{dom} F := \{x \in \mathbb{R}^n : F(x) < +\infty\}$, and its *infimum*

$$\inf F := \inf_{x \in \mathbb{R}^n} F(x) = \inf_{x \in \operatorname{dom} F} F(x).$$

Whenever this infimum is attained, ie. there is some x such that $F(x) = \inf F$, then it is called a minimum and is denoted by $\min F$. We further define

$$\operatorname{argmin} F := \{x \in \mathbb{R}^n : F(x) = \inf F\}.$$

Additionally, a function F is *lower semi-continuous* if for any $x \in \mathbb{R}^n$,

$$\liminf_{u \to x} F(u) := \min\{t \in \overline{\mathbb{R}} : \exists u_r \to x \text{ with } F(u_r) \to t\} = F(x).$$

Finally, a function F is said to be *proper* is $F(x) < +\infty$ for at least one $x \in \mathbb{R}^n$ and $F(x) > -\infty$ for all $x \in \mathbb{R}^n$. This means that the domain of a proper function is a nonempty set over which F is finite-valued.

1.2.3 (Sub)Gradients of convex functions

This class of functions comes with several interesting properties, for instance $\operatorname{dom} F$ and $\operatorname{argmin} F$ are convex if F is convex, furthermore, every local minimum is a global one. This is captured by the notion of subgradients.

Lemma 1.7 (Rockafellar and Wets 1998, Prop. 8.12)**.** *Consider a convex proper function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \operatorname{dom} F$. Then,*

$$\partial F(x) = \{v : F(u) \geq F(x) + \langle v, u - x\rangle \text{ for all } u \in \mathbb{R}^n\} \neq \emptyset \tag{1.4}$$

and $0 \in \partial F(x)$ if and only if $x \in \operatorname{argmin} F$.

An important point is that $u \mapsto F(x)+\langle v, u-x\rangle$ provides a linear under-approximation of the whole function F.

When F is differentiable, then $\partial F(x) = \{\nabla F(x)\}$ and convexity can be seen directly as a property on the gradient mapping.

Theorem 1.8 (Bauschke and Combettes 2011, Prop. 17.10)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function with open domain.*[4] *Suppose that f is differentiable on $\operatorname{dom} f$. Then the following are equivalent:*

i) f is convex;

ii) $f(u) \geq f(x) + \langle \nabla f(x), u - x\rangle$ for all $x, u \in \operatorname{dom} f$;

iii) $\langle \nabla f(x) - \nabla f(u), x - u\rangle \geq 0$ for all $x, u \in \operatorname{dom} f$, ie. ∇f is monotone.

Furthermore, if f is twice differentiable on $\operatorname{dom} f$, any of the above is equivalent to

iv) $\langle u, \nabla^2 f(x) u\rangle \geq 0$ for all $x, u \in \operatorname{dom} f$, ie. $\nabla^2 f$ is positive semi-definite.

[4] *typically here,* $\operatorname{dom} f = \mathbb{R}^n$.

1.2.4 Optimality conditions for convex functions on convex sets

Let us consider the problem of minimizing a convex function F over a convex set C. The problem consists in finding $x^\star \in C$ such that $F(x^\star) \leq F(x)$ for all $x \in C$, we note this problem

$$x^\star \in \operatorname{argmin}_C F \Leftrightarrow x^\star \text{ is a solution of } \inf_{x \in C} F(x)$$

We directly note that if C is empty, the problem is impossible[5] and if C is open it may be impossible to find a solution. Hence, we will restrict our analysis to closed nonempty convex sets as before.

[5] *infeasible* in the optimization language.

The *constrained* variant of Fermat's rule that links the gradient of the function with local minimas writes as follows.

Theorem 1.9 (Rockafellar and Wets 1998, Th. 6.12, 8.15)**.** *Consider a proper lower-semicontinuous convex function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a convex set C. Then, $x \in \operatorname{argmin}_C F$ if and only if $x \in C$ and $0 \in \partial F(x) + N_C(x)$ or,[6] equivalently,*

$$\langle y - x, v \rangle \geq 0$$

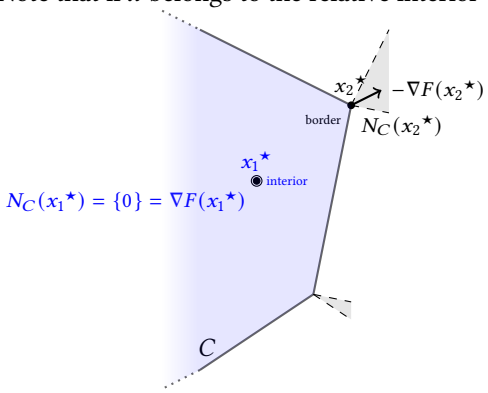
for any $v \in \partial F(x)$ and all $y \in C$.
In particular, if F is differentiable, $0 \in \nabla F(x) + N_C(x)$ means that

$$\langle y - x, \nabla F(x) \rangle \geq 0$$

for all $y \in C$.

[6] *The* normal cone *of a convex set C at a point $x \in C$ is defined as the set $N_C(x) \coloneqq \{u : \langle y - x, u \rangle \leq 0 \text{ for all } y \in C\}$.*

Note that if x belongs to the relative interior of C, then $N_C(x) = \{0\}$.



1.2.5 Strict & strong convexity

Strict convexity is simply convexity but when every inequality is replaced with a *strict inequality*: a function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strictly convex if and only if for any $x, u \in C$, $F((1-\alpha)x + \alpha u) < (1-\alpha)F(x) + \alpha F(u)$ for any $\alpha \in (0, 1)$. All results above then hold with strict inequalities.

Lemma 1.10. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strictly convex lower semi-continuous proper function and C a convex set, then F has at most one minimizer on C. In particular, F has at most one minimizer on $\mathbb{R}^n$.*

Strict convexity can be observed mathematically and from that we can ensure the uniqueness of solutions. However, it is almost impossible to exploit numerically since it

only grants us a strict inequality and not an exploitable knowledge about the function's local behavior. For this, we need a stronger condition: strong convexity. While convexity provides affine lower bounds, strongly convex functions have quadratic lower-bounds enable to get a better control that may have a great impact on the convergence of optimization methods.

Definition 1.11. For some $\mu > 0$, a function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is μ-strongly convex if and only if $F - \frac{1}{2}\mu\|\cdot\|^2$ is convex.

Using the fact that $\tilde{F} \coloneqq F - \frac{1}{2}\mu\|\cdot\|^2$ is convex and verifies $\partial\tilde{F} = \partial F - \mu\cdot$, we get that for any $x \in \mathbb{R}^n$ and any $v \in \partial F(x)$

$$F(u) \geq F(x) + \langle v, u - x\rangle + \frac{\mu}{2}\|u - x\|^2 \text{ for all } u \in \mathbb{R}^n \tag{1.5}$$

which directly implies that a strongly convex function has at most one minimizer by taking x such that $0 \in \partial F(x)$. The following lemma then adds the existence (see (Bauschke and Combettes, 2011, Chap. 11.4) for a more general take).

Lemma 1.12. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a lower semi-continuous proper strongly convex function and C a convex set, then F has exactly one minimizer on C. In particular, F has exactly one minimizer on $\mathbb{R}^n$.*

Proof. ([⋆]) Let us consider the case where $C = \mathbb{R}^n$, the other cases can be deduced easily. From (1.5), we get that for all $u \in \mathbb{R}^n$,

$$\begin{aligned} F(u) &\geq F(x) + \frac{\mu}{2}\|x\|^2 - \langle v, x\rangle + \langle v + \mu x, u\rangle + \frac{\mu}{2}\|u\|^2 \\ &\geq F(x) + \frac{\mu}{2}\|x\|^2 - \langle v, x\rangle - \|v + \mu x\|\|u\| + \frac{\mu}{2}\|u\|^2 \end{aligned}$$

hence $F(u)/\|u\| \to +\infty$ when $\|u\| \to +\infty$, ie. F is supercoercive. Thus, this means that for any t, the level set $\{x : F(x) \leq t\}$ is bounded (this is direct by contradiction, see (Bauschke and Combettes, 2011, Chap. 11.11)). Since F is proper, we can take t sufficiently large so that the corresponding level set is non-empty and bounded. Finally, since F is lower semi-continuous, applying Lemma 1.1 to this compact set gives us the existence of a minimal value, which is unique from the quadratic lower bound expressed in (1.5). □

If a differentiable function is strongly convex, we have the following characterizations.

Theorem 1.13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function with open domain. Suppose that f is differentiable on* $\operatorname{dom} f$*. Then the following are equivalent:*

i) f is μ-strongly convex;

ii) $f(u) \geq f(x) + \langle \nabla f(x), u - x\rangle + \frac{\mu}{2}\|u - x\|^2$ for all $x, u \in \operatorname{dom} f$;

iii) $\langle \nabla f(x) - \nabla f(u), x - u\rangle \geq \mu\|u - x\|^2$ for all $x, u \in \operatorname{dom} f$, ie. ∇f is monotone.

Furthermore, if f is twice differentiable on $\operatorname{dom} f$*, any of the above is equivalent to*

iv) $\langle u, \nabla^2 f(x)u\rangle \geq \mu\|u\|^2$ for all $x, u \in \operatorname{dom} f$, ie. $\nabla^2 f$ is positive definite.

CHAPTER 2 Gradient and Subgradient methods

Gradient methods are the most simple optimization algorithm. They are built upon, the idea that differentiating the function tells you in which direction to go to minimize the function value. However, gradient heavily rely on smoothness, and things can go awry in other situations.

2.1 Smoothness and gradient descent

The Gradient descent algorithm on a differentiable function function f consists in taking $x_0 \in \mathbb{R}^n$ and iterating

$$x_{k+1} = x_k - \gamma \nabla f(x_k) \qquad \text{(Gradient descent)}$$

for some $\gamma > 0$.

2.1.1 Smoothness

There is slight discrepancy in the literature concerning the notion of smoothness for functions. In (Rockafellar and Wets, 1998), it is used for continuously differentiable functions, in Riemannian analysis it often refers to C^∞ function, while in numerical optimization and machine learning (see eg. (Bubeck et al., 2015)), it is used for functions with Lipschitz-continuous gradients. We will adopt the latter viewpoint. The reason for this is that it allows us to have a quadratic upper approximation of our function, obtained directly from the fundamental theorem of calculus. This is the crucial point for the use of gradient methods.

Definition 2.1. We say that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is L-smooth if it has a L-Lipschitz continuous gradient, ie. if

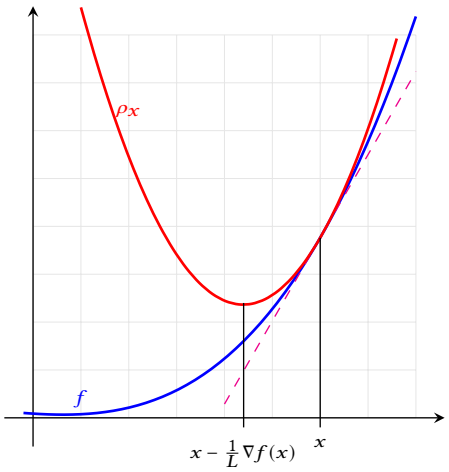
$$\|\nabla f(x) - \nabla f(u)\| \leq L\|x - u\| \text{ for all } x, u \in \mathbb{R}^n.$$

From this property, we can derive this highly important lemma.

Lemma 2.2. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with a L-Lipschitz continuous gradient, then for any $x, u \in \mathbb{R}^n$, one has*

$$|f(u) - f(x) - \langle \nabla f(x), u - x \rangle| \leq \frac{L}{2}\|x - u\|^2.$$

Thus, if we fix a point x, the function $\rho_x : u \mapsto f(x) + \langle \nabla f(x), u - x \rangle + \frac{L}{2}\|u - x\|^2$ is quadratic in its argument and majorizes f, that is to say $\rho_x(u) \geq f(u)$ for any u. Furthermore, the minimum of ρ_x is attained at $x^\star = x - \frac{1}{L}\nabla f(x)$.



Such a quadratic approximation can be leveraged using gradients steps, ie. taking

$$u = x - \gamma \nabla f(x)$$

for some $\gamma > 0$. Indeed, in that case, Lemma 2.2 gives us

$$f(u) \leq f(x) - \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x - u\|^2 = f(x) - \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla f(x)\|^2. \tag{2.1}$$

2.1.2 Gradient algorithm for convex functions

When f is L-smooth and convex, we can guarantee convergence and a $O(1/k)$ rate.

Theorem 2.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex L-smooth function. Then, the iterates (x_k) generated by (Gradient descent) with $\gamma = 1/L$ satisfy:*

- *(convergence) $x_k \to x^\star$ for some minimizer $x^\star$ of f;*[7]
- *(rate) $f(x_k) - f(x^\star) \leq \dfrac{2L\|x_0 - x^\star\|^2}{k}$ for any minimizer $x^\star$ of f.*

[7] *ie. a point such that $\nabla f(x^\star) = 0$.*

In the above theorem, any $\gamma \in (0, 1/L)$ actually works for the convergence and gets a similar complexity but $\gamma = 1/L$ is the optimal value in terms of rate.

Remark 2.4 (Lower bound). This is not the fastest way to minimize a convex smooth function. Actually, one can show that the fastest attainable rate for this class of functions is $O(1/k^2)$; see (Bubeck et al., 2015, Th. 3.14). This complexity is attained by Nesterov's fast gradient method (Nesterov, 1983). This method accelerates gradient descent by adding an "inertial" step:

$$\begin{aligned} y_{k+1} &= x_k - \gamma \nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \alpha_{k+1}(y_{k+1} - y_k) \end{aligned} \qquad \text{(Fast Gradient descent)}$$

where $\gamma \in (0, 1/L)$ and $\alpha_{k+1} = (k+2)/(k+3)$.[8] ◂

[8] Actually, the choice for α_{k+1} is a bit more complicated but this variant grants the same rate.

2.1.3 Gradient algorithm for strongly convex functions

Now, if the function is additionally strongly convex, the quadratic lower bounds grants us a better rate.

Theorem 2.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex L-smooth function. Then, the iterates (x_k) generated by* (Gradient descent) *with $\gamma = \frac{2}{\mu+L}$ satisfy:*

- *(convergence) $x_k \to x^\star$ for the minimizer $x^\star$ of f;*[9]
- *(rate) $f(x_k) - f(x^\star) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|x_0 - x^\star\|^2$ where $\kappa = \frac{L}{\mu} \geq 1$.*

[9] *unique by strong convexity*

In the above theorem, any $\gamma \in (0, 2/(\mu+L)]$ actually works for the convergence and gets a similar complexity but $\gamma = 2/(\mu+L)$ is the optimal value in terms of rate.

We note here that the term $\kappa = \frac{L}{\mu} \geq 1$ appears in the rate, this number is generally called the *conditioning* of the number by analogy with matrices and linear systems.

Finally, the obtained rate is again not optimal for this class of functions, the optimal rate being $O\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\right)$, again attained by a modified version of (Fast Gradient descent).

2.1.4 Projected Gradient algorithm

Now let us consider the problem of minimizing a smooth convex function f over a nonempty closed convex set C. Thanks to the ability to project onto C, we can easily define a projected gradient method:

$$x_{k+1} = \text{proj}_C\left(x_k - \gamma \nabla f(x_k)\right) \qquad \text{(Projected gradient descent)}$$

for some initialization $x_0 \in \mathbb{R}^n$ and stepsize $\gamma > 0$.

This algorithm has similar guarantees as gradient descent.

Theorem 2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex L-smooth function. Then, the iterates (x_k) generated by* (Gradient descent) *with $\gamma = 1/L$ belong to C and satisfy:*

- *(convergence) $x_k \to x^\star$ for some minimizer $x^\star$ of f on C;*[10]
- *(rate) $f(x_k) - f(x^\star) \leq \dfrac{3L\|x_0 - x^\star\|^2 + f(x_0) - f(x^\star)}{k+1}$ for any minimizer $x^\star$ of f on C.*

[10] *ie. a point such that $-\nabla f(x^\star) \in N_C(x^\star)$, ie. $\langle y - x^\star, \nabla F(x^\star)\rangle \geq 0$ for all $y \in C$.*

2.2 Nonsmooth (sub)gradient descent

If our function is nonsmooth (which is one of the core topics of the course), things change quite a lot. In this section, we see what happens when for (sub)gradient-based algorithm when our function is i) non-differentiable; and ii) differentiable but nonsmooth.

2.2.1 Non-differentiability & subgradient descent

A direct method to minimize a convex non-differentiable function g is to mimic the gradient method and to do subgradient descent:

$$x_{k+1} = x_k - \gamma_k v_k \text{ with } v_k \in \partial g(x_k) \qquad \text{(Subgradient descent)}$$

Here, a fixed stepsize is not always possible. For instance, take $g = |\cdot|$, then (x_k) will oscillate around 0 for any $\gamma > 0$.

In fact, we have the following result.

Theorem 2.7. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semi-continuous convex function with a minimizer $x^\star$. Assume that $\|v\| \leq M$ for any $x \in \operatorname{dom} g$ and any $v \in \partial g(x)$. Then, the Subgradient descent algorithm started with x_0 generates iterates that verify:*

a) for a constant stepsize $\gamma_k = \gamma$,

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2k\gamma} + \frac{\gamma M^2}{2}.$$

b) for a stepsize sequence verifying $\sum_{k=0}^\infty \gamma_k = +\infty$ and $\sum_{k=0}^\infty \gamma_k^2 < +\infty$,

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \xrightarrow{k\to\infty} 0.$$

Proof. We assume here that g has a minimizer, say $x^\star$. Then, for any k,

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|x_k - \gamma_k v_k - x^\star\|^2 \\ &= \|x_k - x^\star\|^2 - 2\gamma_k \langle v_k; x_k - x^\star\rangle + \gamma_k^2 \|v_k\|^2\end{aligned}$$

Now, since $v_k \in \partial g(x_k)$, Lemma 1.7 with $u = x^\star$ tells us that $g(x^\star) \geq \gamma_k + \langle v_k, x^\star - x_k\rangle$ and thus

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &\leq \|x_k - x^\star\|^2 - 2\gamma_k (g(x_k) - g(x^\star)) + \gamma_k^2 \|v_k\|^2 \\ &\leq \|x_0 - x^\star\|^2 - 2\sum_{\ell=0}^k \gamma_\ell (g(x_\ell) - g(x^\star)) + \sum_{\ell=0}^k \gamma_\ell^2 \|v_\ell\|^2.\end{aligned}$$

This enables us to get that

$$\frac{\sum_{\ell=0}^k \gamma_\ell (g(x_\ell) - g(x^\star))}{\sum_{\ell=0}^k \gamma_\ell} \leq \frac{\|x_0 - x^\star\|^2 + \sum_{\ell=0}^k \gamma_\ell^2 \|v_\ell\|^2}{2\sum_{\ell=0}^k \gamma_\ell}.$$

First, we notice that by convexity,

$$\min_{\ell \leq k} g(x_\ell) - g(x^\star) \leq g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\sum_{\ell=0}^k \gamma_\ell (g(x_\ell) - g(x^\star))}{\sum_{\ell=0}^k \gamma_\ell}. \tag{2.2}$$

As for the right hand size:
(a) if $\gamma_k = \gamma$, then

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2k\gamma} + \frac{\gamma M^2}{2}.$$

if $\sum_{k=0}^\infty \gamma_k = +\infty$ and $\sum_{k=0}^\infty \gamma_k^2 < +\infty$, then

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\|x_0 - x^\star\|^2 + M^2 \sum_{\ell=0}^\infty \gamma_\ell^2}{2\sum_{\ell=0}^k \gamma_\ell}$$

and the RHS's numerator is finite while the denominator is going to infinity as $k \to \infty$, the whole term thus goes to 0. □

Note that the result above also holds for $\min_{\ell \leq k} g(x_\ell) - g(x^\star)$ by (2.2). However, since the stepsize is decreasing, this limits the rate and the iterates convergence is out of reach. Nevertheless, its rate in $O(1/\sqrt{k})$ is optimal on this class of functions.

It is also possible to add a projection to a convex set, the proof only changes in the first line where the non-expansiveness of the projection has to be used. More precisely, the algorithm

$$x_{k+1} = \mathrm{proj}_C\,(x_k - \gamma_k\, v_k) \;\text{ with } v_k \in \partial g(x_k) \qquad \text{(Projected Subgradient descent)}$$

verifies the following properties.

Theorem 2.8. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semi-continuous convex function and let C be a closed convex set. Assume that $\|v\| \leq M$ for any $x \in C$ and any $v \in \partial g(x)$. Then, g has a minimizer $x^\star$ in C and the Projected Subgradient descent algorithm started with $x_0 \in C$ generates iterates that verify:*

a) for a constant stepsize $\gamma_k = \gamma$,

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2k\gamma} + \frac{\gamma M^2}{2}.$$

b) for a stepsize sequence verifying $\sum_{k=0}^\infty \gamma_k = +\infty$ and $\sum_{k=0}^\infty \gamma_k^2 < +\infty$,

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \xrightarrow{k\to\infty} 0.$$

2.2.2 Nonsmoothness & gradient descent

When you function is differentiable, you still have that

$$g(x + t\nabla g(x)) = g(x) - t\|\nabla g(x)\|^2 + o(t\|\nabla g(x)\|)$$

which implies that

$$\frac{g(x + t\nabla g(x)) - g(x)}{t} = -\|\nabla g(x)\|^2 + o(1)$$

and thus you can still find a small enough step t that will decrease you functional value, for instance using line-search methods. Unfortunately, to translate this property to some convergence result smoothness is needed.

Thus, we have two paths to overcome this problem:

a) Changing our algorithm. Taking a look at a gradient step, we notice that

$$\begin{aligned} y = x - \gamma\nabla g(x) &\Leftrightarrow \nabla g(x) + \frac{y - x}{\gamma} = 0 \\ &\Leftrightarrow y = \mathrm{argmin}_u \left\{\langle g(x); u\rangle + \frac{1}{2\gamma}\|u - x\|^2\right\} \\ &\Leftrightarrow y = \mathrm{argmin}_u \left\{g(x) + \langle g(x); u - x\rangle + \frac{1}{2\gamma}\|u - x\|^2\right\} \end{aligned}$$

and if smoothness is lacking, maybe changing the first order approximation can help. This is what we will do in Chapter 3.

b) Changing our definition of smoothness. The smoothness property:

$$g(u) \leq g(x) + \langle \nabla g(x), u - x \rangle + \frac{L}{2}\|x - u\|^2$$

can be rewritten as

$$\left(\frac{L}{2}\|u\|^2 - g(u)\right) - \left(\frac{L}{2}\|x\|^2 - g(x)\right) \leq \langle Lx - \nabla g(x), u - x \rangle$$

which is equivalent to saying that $x \mapsto \frac{L}{2}\|x\|^2 - g(x)$ is convex.
This indicates that smoothness is intricately linked with the squared Euclidean norm. To deal with functions that are not smooth, a good idea is thus to change how we measure distances.

2.3 Non-Euclidean gradient descent

In this section, we focus on methods in which the Euclidean distance is replaced by a Bregman divergence. We thus study how gradient methods translate in this case.

Our template problem for this section will be

$$\min_{x \in C} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function and $C \subset \mathbb{R}^n$ is a closed convex set with non-empty interior.

2.3.1 Bregman divergences

The core ingredient for defining divergences is the notion of *Bregman regularizer* – or *distance-generating function* (DGF) – which we define below as follows:

Definition 2.9 (Bregman regularizers)**.** A proper lower semi-continuous strictly convex function $h \colon \mathcal{X} \to \overline{\mathbb{R}}$ is said to be a *Bregman regularizer* on C if

(i) h is supported on C, i.e., $\operatorname{dom} h = C$.

(ii) the subgradient of h admits a *continuous selection*, i.e., there exists a continuous mapping ∇h such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$.

(iii) h is 1-strongly convex relative to $\|\cdot\|$, i.e., for all $x \in \operatorname{dom} \partial h, u \in \operatorname{dom} h$, we have

$$h(u) \geq h(x) + \langle \nabla h(x), u - x \rangle + \frac{1}{2}\|u - x\|^2.$$

Note that the norm above is no longer necessarily the Euclidean one. An important set is $C_h \coloneqq \operatorname{dom} \partial h$ is called the *prox-domain* of h; we have $\operatorname{ri} C \subseteq C_h \subseteq C$.

From a regularizer, one can define a divergence as follows.

Definition 2.10 (Bregman divergence)**.** From a Bregman regularizer h, we define the associated *Bregman divergence* as

$$D(u, x) = h(u) - h(x) - \langle \nabla h(x), u - x \rangle \qquad \text{for all } x \in C_h,\, u \in C$$

Is is immediate to see that $D(u, x) \geq 1/2\|u - x\|^2 \geq 0$. However, these divergences do not admit exactly developments like for the squared Euclidean norm but they verify the following inequality.

Lemma 2.11 (3-point identity)**.** *Let h be a Bregman regularizer on C. For all $u \in C$ and all $x, p \in C_h$, we have:*

$$D(u,p) = D(u,x) + D(x,p) + \langle \nabla h(p) - \nabla h(x), x - u \rangle$$

Example 2.12 (Euclidean regularization). Perhaps the most widely used DGF is the quadratic regularizer $h(x) = x^2/2$ for $x \in C$. Concretely, taking $C = [0, \infty)$ and noting that $h'(x) = x$, we have:

a) Prox-domain: $C_h = C$

b) Bregman divergence: $D(p,x) = (p-x)^2/2$

Example 2.13 (Entropic regularization). Another popular choice when $C = [0, \infty)$ is the entropic regularizer $h(x) = x \log x$. In this case, we have $h'(x) = 1 + \log x$, which gives the following:

a) Prox-domain: $C_h = \operatorname{ri} C = (0, \infty)$

b) Bregman divergence: $D(p,x) = p \log(p/x) + x - p$

which is the Kullback-Liebler divergence.

2.3.2 Gradient descent with Bregman divergences

First, let us see precisely where the distance can appear when we are using projected gradient descent. Recall that since C is a closed convex set

$$\begin{aligned}
& y = \operatorname{proj}_C (x - \gamma \nabla g(x)) \\
\Leftrightarrow & y = \operatorname{argmin}_{u \in C} \left\{ \frac{1}{2} \| u - (x - \gamma \nabla g(x)) \|_2^2 \right\} \text{ (by Lemma 1.3)} \\
\Leftrightarrow & y = \operatorname{argmin}_{u \in C} \left\{ \langle -\gamma \nabla g(x); x - u \rangle + \frac{1}{2} \| u - x \|_2^2 \right\}
\end{aligned}$$

We see that a Euclidean distance between u and x appears (coming from the projection operator and the fact that in order to isolate the gradient in a scalar product, developing the norm is needed). Naturally, we can replace this distance by the Bregman divergence introduced above.

Definition 2.14 (Bregman proximal mapping)**.** From a Bregman regularizer h, we define the induced *Bregman proximal mapping* as

$$P_x(y) = \operatorname{argmin}_{u \in C} \{ \langle y, x - u \rangle + D(u,x) \} \qquad \text{for all } x \in C_h, y \in \mathcal{Y}.$$

When the Bregman divergence amounts to the Euclidean distance, we obtain the vanilla projected gradient for $y = -\gamma \nabla g(x)$.

However, the existence of this mapping is not for granted, it strongly relies on the properties of the Bregman regularizer.

Lemma 2.15. *Let h be a Bregman regularizer, then for all $x \in C_h$, $y \in \mathcal{Y}$, the proximal mapping $P_x(y)$ is uniquely defined and belongs to C_h.*

Proof. First, let develop the divergence:

$$\begin{aligned}
P_x(y) &= \operatorname{argmin}_{u \in C} \{ \langle y, x - u \rangle + D(u,x) \} \\
&= \operatorname{argmin}_{u \in C} \{ \langle y, x - u \rangle + h(u) - h(x) - \langle \nabla h(x), u - x \rangle \} \\
&= \operatorname{argmin}_{u \in C} \{ h(u) - \langle y + \nabla h(x), u \rangle \}
\end{aligned}$$

Recalling that h is strongly convex with respect to some norm, the reasoning of Lemma 1.12 can adapted to show that $P_x(y)$ is well defined and unique.

Furthermore, noting $v = y + \nabla h(x)$, we get that

$$h(P_x(y)) - \langle v, P_x(y)\rangle \leq h(u) - \langle v, u\rangle$$

for all $u \in C$. This means that

$$h(u) \geq h(P_x(y)) + \langle v, u - P_x(y)\rangle$$

for all u, which implies that v is a subgradient of h at $P_x(y)$. Thus, $P_x(y) \in \operatorname{dom} \partial h = C_h$. □

Now, we can state and analyze our (sub)gradient descent with Bregman divergence, also called *Mirror descent*.

$$x_{k+1} = P_{x_k}(-\gamma_k \nabla g(x_k))) \qquad \text{(Mirror Descent)}$$

The analysis can be carried out in a similar way to the projected subgradient descent.

Theorem 2.16. *Let g be a convex proper lower-semicontinuous function and let C be a closed convex set. Take h a Bregman regularizer on C. Then, if g is differentiable on* $\operatorname{dom} \partial h$ *and L Lipschitz-continuous on with respect to $\|\cdot\|$, the iterates of Mirror Descent verify*

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(u) \leq \frac{D(u, x_0) + \sum_{\ell=0}^k \gamma_\ell^2 L^2}{2\sum_{\ell=0}^k \gamma_\ell}$$

for any $u \in C$.

Proof. From the optimality conditions of

$$\begin{aligned} x_{k+1} &= P_{x_k}(-\gamma_k \nabla g(x_k))) \\ &= \operatorname{argmin}_{u\in C}\{h(u) - \langle -\gamma_k \nabla g(x_k) + \nabla h(x_k), u\rangle\}, \end{aligned}$$

we get that

$$\langle \nabla h(x_{k+1}) + \gamma_k \nabla g(x_k) - \nabla h(x_k), u - x_{k+1}\rangle \geq 0$$

for all $u \in C$. Thus,

$$\begin{aligned} \gamma_k \langle \nabla g(x_k), x_k - u\rangle &\leq \langle \nabla h(x_{k+1}) + \gamma_k \nabla g(x_k) - \nabla h(x_k), u - x_{k+1}\rangle + \gamma_k \langle \nabla g(x_k), x_k - u\rangle \\ &= \gamma_k \langle \nabla g(x_k); x_k - x_{k+1}\rangle \\ &\quad + \langle \nabla h(x_{k+1}) - \nabla h(x_k), u - x_{k+1}\rangle \\ &= \gamma_k \langle \nabla g(x_k); x_k - x_{k+1}\rangle \\ &\quad + D(u, x_k) - D(u, x_{k+1}) - D(x_{k+1}, x_k) \\ &\leq \gamma_k L\|x_k - x_{k+1}\| - \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\quad + D(u, x_k) - D(u, x_{k+1}) \\ &\leq \max_s\{\gamma_k L s - \frac{1}{2}s^2\} + D(u, x_k) - D(u, x_{k+1}) \\ &= \frac{\gamma_k^2 L^2}{2} + D(u, x_k) - D(u, x_{k+1}) \end{aligned}$$

where we successively used the three point identity (Lemma 2.11), the strong convexity of the Bregman regularizer, and the Lipschitz continuity of g.

By summing for $\ell = 0, .., k$, we get

$$g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(u) \leq \frac{\sum_{\ell=0}^k \gamma_\ell (g(x_\ell) - g(u))}{\sum_{\ell=0}^k \gamma_\ell} \leq \frac{D(u, x_0) + \sum_{\ell=0}^k \gamma_\ell^2 L^2}{2 \sum_{\ell=0}^k \gamma_\ell}.$$

and the rest of the proof is the same as for subgradient descent (Theorem 2.7). □

2.3.3 *Application example: minimization over the simplex*

The simplex is the set $C = \Delta_n = \{x \in \mathbb{R}^n : x \geq 0 \text{ and } \sum_{i=1}^n x_i = 1\}$, for which a useful regularizer is

$$h(x) = \sum_{i=1}^n x_i \log(x_i)$$

which gives the KL divergence:

$$D(u, x) = \sum_{i=1}^n u_i \log\left(\frac{u_i}{x_i}\right) - \sum_{i=1}^n (u_i - x_i).$$

We can verify that $\operatorname{dom} h = \Delta_n$, h is differentiable except when one of the coordinates is null, ie. $C_h = \operatorname{dom} \partial h = \operatorname{ri} \Delta_n$. Finally, h is 1-strongly convex with respect to the 1-norm $\|\cdot\|_1$, this is known as Pinsker's inequality.

Then, $p = P_x(y)$ is easily computed since

$$\begin{aligned}
& p = P_x(y) = \operatorname{argmin}_{u \in C} \{h(u) - \langle y + \nabla h(x), u\rangle\} \\
\Leftrightarrow & \sum_{i=1}^n \left[(1 + \log(p_i)) - y_i - (1 + \log(x_i))\right] [s_i - x_i] \geq 0 \text{ for all } s \in \Delta_n \\
\Leftrightarrow & \sum_{i=1}^n \left[\log(p_i) - \log(x_i \exp(y_i))\right] [s_i - x_i] \geq 0 \text{ for all } s \in \Delta_n
\end{aligned}$$

It is possible to take $c > 0$ and set $p_i = c.x_i \exp(y_i))$ for all i $\sum_{i=1}^n s_i - x_i = 0$. Since $x_i \geq 0$, the only c guaranteeing that $p \in \Delta_n$ is the one that verifies

$$\sum_{i=1}^n p_i = \sum_{i=1}^n c.x_i \exp(y_i) = 1$$

i.e., $c = \frac{1}{\sum_{i=1}^n x_i \exp(y_i))}$. Finally, we get that

$$p_i = P_x(y)_i = \frac{x_i \exp(y_i)}{\sum_{i=1}^n x_i \exp(y_i))}.$$

A gradient step on the simplex with an entropic regularizer is called the multiplicative weights updates and writes

$$y_i = \frac{x_i \exp(-\gamma \nabla_i g(x))}{\sum_{j=1}^n x_j \exp(-\gamma \nabla_j g(x))}.$$

2.3.4 What do we gain compared to projected (sub)gradient descent?

Easier projections and constants

For the simplex above, we completely avoided the projection step (which was handled by the renormalization). In addition, the radius term in the error estimate is in $D(u, x_0)$ instead of $\|u - x_0\|^2$. For the simplex, the former is in $\log(n)$ while the latter is in n.

Smoothness with respect to the regularizer

Recall that if

$$\begin{aligned} x_{k+1} &= P_{x_k}(-\gamma_k \nabla g(x_k))) \\ &= \operatorname{argmin}_{u \in C}\{h(u) - \langle -\gamma_k \nabla g(x_k) + \nabla h(x_k), u\rangle\}, \end{aligned}$$

we get that

$$\begin{aligned} \gamma_k(g(x_k) - g(u)) &\leq \gamma_k \langle \nabla g(x_k), x_k - u\rangle \\ &\leq \langle \nabla h(x_{k+1}) + \gamma_k \nabla g(x_k) - \nabla h(x_k), u - x_{k+1}\rangle + \gamma_k \langle \nabla g(x_k), x_k - u\rangle \\ &= \gamma_k \langle \nabla g(x_k); x_k - x_{k+1}\rangle \\ &\quad + \langle \nabla h(x_{k+1}) - \nabla h(x_k), u - x_{k+1}\rangle \\ &= \gamma_k \langle \nabla g(x_k); x_k - x_{k+1}\rangle - D(x_{k+1}, x_k) \\ &\quad + D(u, x_k) - D(u, x_{k+1}) \end{aligned}$$

thus if we have some relative smoothness inequality:

$$g(u) \leq g(x) + \langle \nabla g(x), u - x\rangle + LD(u, x)$$

then

$$\langle \nabla g(x_k); x_k - x_{k+1}\rangle \leq g(x_k) - g(x_{k+1}) + LD(x_{k+1}, x_k)$$

which gives

$$\begin{aligned} \gamma_k(g(x_k) - g(u)) \leq \gamma_k(g(x_k) - g(x_{k+1})) - (1 - \gamma_k L)D(x_{k+1}, x_k) \\ + D(u, x_k) - D(u, x_{k+1}) \end{aligned}$$

which implies i) that $(g(x_k))$ is non-increasing (take $u = x_k$); and ii) better convergence rates can be obtained (typically in $O(1/k)$). For more details, see (Bauschke et al., 2017).

CHAPTER 3 Proximal and Bundle methods

Going beyond first order can be highly beneficial in nonsmooth optimization in order not to rely on the rather loose local information brought by the subgradient.

In the previous section, we investigated first-order methods that could be seen as iterations of the type

$$\begin{aligned} & x_{k+1} = x_k - \gamma\, v_k \text{ with } v_k \in \partial F(x_k) \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ \frac{1}{2}\|u - (x_k - \gamma\, v_k)\|_2^2 \right\} \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ \langle -\gamma\, v_k; x_k - u\rangle + \frac{1}{2}\|u - x_k\|_2^2 \right\} \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ F(x_k) + \langle v_k; u - x_k\rangle + \frac{1}{2\gamma}\|u - x_k\|_2^2 \right\}. \end{aligned}$$

Recalling that by definition (see (1.4))

$$\partial F(x_k) = \{v : F(u) \geq F(x_k) + \langle v, u - x_k\rangle \text{ for all } u \in \mathbb{R}^n\},$$

a subgradient step can be seen as:

$$x_{k+1} = \operatorname{argmin}_u \left\{ \underbrace{F(x_k) + \langle v_k; u - x_k\rangle}_{\text{(a)}} + \underbrace{\frac{1}{2\gamma}\|u - x_k\|_2^2}_{\text{(b)}} \right\}.$$

where:

(a) is a linear/first-order model that under-approximates F;

(b) is a quadratic recall/regularization/stabilization term controlled by γ.

In this chapter, we will investigate algorithms that minimize stabilized approximations of the function.

3.1 The Proximity Operator

A central tool to tackle nonsmooth functions is the *proximity operator*, introduced by (Moreau, 1965), and denoted $\mathbf{prox}_{\gamma F}$ for a step-size $\gamma > 0$ and a nonsmooth function

$F : \mathbb{R}^n \to \overline{\mathbb{R}}$; it is defined as the set-valued mapping

$$\mathbf{prox}_{\gamma F}(y) \coloneqq \operatorname{argmin}_{u \in \mathbb{R}^n} \underbrace{\left\{ F(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}}_{\coloneqq \rho_y(u)}.$$

In the same flavor as for the gradient step, if one takes a proximal step, ie.

$$x = \mathbf{prox}_{\gamma F}(y)$$

for some $\gamma > 0$, the definition directly gives us

$$F(x) \leq F(y) - \frac{1}{2\gamma}\|x - y\|^2 \tag{3.1}$$

which mirrors (2.1) (the descent inequality of a gradient step on a smooth function) but for a nonsmooth function.[11]

With this respect, the proximity operator provides a alternative to the use of subgradients or nonsmooth gradients since they are not able to provide descent inequalities such as (2.1) and (3.1). However, this comes with the cost of having to solve a minimization subproblem, which in turn question about the existence and uniqueness of the subproblem solutions.

[11] Actually, this link can be made formal since a proximal step is equivalent to a gradient step on the *Moreau envelope* defined for all $y \in \mathbb{R}^n$ as $e_\gamma F(y) = \inf_{u \in \mathbb{R}^n} \rho_y(u)$ (Moreau, 1965; Yosida, 1988).

3.1.1 Properties

First, for convex functions the proximity operator exists and is unique.

Theorem 3.1. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $\mathbf{prox}_{\gamma F}(y)$ is a singleton for any $\gamma > 0$ and any $y \in \mathbb{R}^n$.*

Proof. Since F is convex, ρ_y is $\frac{1}{\gamma}$-strongly convex. Then, Lemma 1.12 guarantees the existence and uniqueness of the minimizers of $\rho_y(u)$ for any u, which means that $\mathbf{prox}_{\gamma F}(y)$ is well-defined and unique. □

In addition, we have the following identity which notably shows that

$$x = \mathbf{prox}_{\gamma F}(y)$$

is equivalent to having

$$x = y - \gamma v \text{ with } v \in F(x)$$

and thus the proximal operator can be seen as an *implicit* subgradient descent step.

Proposition 3.2. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then the following propositions are equivalent:*

i) $x = \mathbf{prox}_{\gamma F}(y)$;

ii) $(y - x)/\gamma \in \partial F(x)$;

iii) $F(u) \geq F(x) + \langle (y - x)/\gamma, u - x \rangle$ *for any* $u \in \mathbb{R}^n$.

Proof. This follow directly from Fermat's rule and the definition of a convex subgradient. □

The above proposition also enables us to show that the proximity operator is (firmly) non-expansive.

3.1.2 Convergence of the proximal point algorithm

Now, let us investigate the proximal point algorithm:

$$x_{k+1} = \mathbf{prox}_{\gamma F}(x_k) \qquad \text{(Proximal Point)}$$

The first thing to notice is that the fixed points of this algorithm correspond to the minimizers of F.

Corollary 3.3. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $x^\star$ is a minimizer of F if and only if $x^\star = \mathbf{prox}_{\gamma F}(x^\star)$ (for any $\gamma > 0$).*

Proof. From Proposition 3.2, we have that $x^\star = \mathbf{prox}_{\gamma F}(x^\star)$ if and only if $0 \in \partial F(x^\star)$ which is equivalent to $x^\star$ being a minimizer of F since it is convex. □

Now, we can analyze the convergence of our proximal point method.

Theorem 3.4. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function. Then, the Proximal Point method with $\gamma > 0$ verifies $F(x_{k+1}) \leq F(x_k)$ and*

- *(convergence) $x_k \to x^\star$ for some minimizer $x^\star$ of F;*
- *(rate) $F(x_k) - F(x^\star) \leq \frac{\|x^\star - x_0\|^2}{2\gamma k}$.*

Proof. We left the iterates convergence proof as an exercise, its reasoning is exactly the same as the one for the gradient algorithm. For the rate, since $x_{k+1} = \mathbf{prox}_{\gamma F}(x_k)$,

$$F(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \leq F(x_k)$$

and thus $F(x_{k+1}) \leq F(x_k)$.

Since $x_{k+1} = \mathbf{prox}_{\gamma F}(x_k)$, it is the minimum of the $1/\gamma$-strongly convex function ρ_{x_k},[12] thus

$$F(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \leq F(x^\star) + \frac{1}{2\gamma}\|x^\star - x_k\|^2 - \frac{1}{2\gamma}\|x_{k+1} - x^\star\|^2$$

[12] If $x^\star$ is the minimizer of a μ-strongly convex function F, then $0 \in \partial F(x^\star)$ and (1.5) gives us that $F(x^\star) \leq F(u) - \frac{\mu}{2}\|u - x^\star\|^2$.

and by summing this inequality from $t = 0, .., k-1$, we get

$$\begin{aligned}\sum_{t=0}^{k-1}\big(F(x_{t+1}) - F(x^\star)\big) &\leq \frac{1}{2\gamma}\sum_{t=0}^{k-1}(\|x^\star - x_t\|^2 - \|x_{t+1} - x^\star\|^2) - \sum_{t=0}^{k-1}\frac{1}{2\gamma}\|x_{t+1} - x_t\|^2 \\ &\leq \frac{1}{2\gamma}\|x^\star - x_0\|^2.\end{aligned}$$

Now, since $F(x_{k+1}) \leq F(x_k)$, we get that

$$k\,\big(F(x_k) - F(x^\star)\big) \leq \sum_{t=0}^{k-1}\big(F(x_t) - F(x^\star)\big) \leq \frac{1}{2\gamma}\|x^\star - x_0\|^2$$

which gives the result. □

3.1.3 Examples of closed form expressions

Example 3.5 (Squared norm). For $F(x) = \frac{1}{2}\|x\|^2$, the proximity operator can be computed explicitly. Since $\rho_y : u \mapsto F(u) + \frac{1}{2\gamma}\|u - y\|^2$ is strongly convex, there is a unique minimizer x and it verifies $\nabla\rho_y(x) = 0$. Thus $x + \frac{1}{\gamma}(x - y) = 0$ which implies $x = y/(1+\gamma)$:

$$\mathbf{prox}_{\gamma\frac{1}{2}\|\cdot\|^2}(y) = \frac{y}{1+\gamma}.$$

Example 3.6 (Projection). In optimization, it is useful to define the *indicator* of set $C \subset \mathbb{R}^n$ as the function $\iota_C : \mathbb{R}^n \to \overline{\mathbb{R}}$ such that[13]

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{else} \end{cases}.$$

[13] This is different from the indicator I_A in probability which is equal to 1 if A is true and 0 elsewhere.

By noticing that the inner minimization in $\mathbb{R}^n$ amounts to minimizing only over C since otherwise the inner value is $+\infty$, this exactly gives the definition of the projection operator. Thus, for $C \subset \mathbb{R}^n$ a non-empty closed convex set and any $\gamma > 0$,

$$\mathbf{prox}_{\gamma\iota_C}(y) = \mathrm{proj}_C(y).$$

Note that the stepsize does not play any role here.

Example 3.7 (Absolute value). The proximity operator of the absolute value admits a closed form expression: for $y \in \mathbb{R}$ and $\gamma > 0$,

$$\mathbf{prox}_{\gamma|\cdot|}(y) = \begin{cases} y + \gamma & \text{if } y < -\gamma \\ 0 & \text{if } -\gamma \le y \le \gamma \\ y - \gamma & \text{if } y > \gamma \end{cases}$$

[14] By coordinates, or blocks of coordinates.

A very useful calculus rule for the proximity operator is that if F is separable:[14]

$$F(x_1, x_2, .., x_m) = \sum_{i=1}^{m} F_i(x_i),$$

then the proximity operator of F can be obtained from those of the (F_i):

$$\mathbf{prox}_{\gamma F}(y_1, y_2, .., y_m) = \begin{bmatrix} \mathbf{prox}_{\gamma F_1}(y_1) \\ \mathbf{prox}_{\gamma F_2}(y_2) \\ \vdots \\ \mathbf{prox}_{\gamma F_m}(y_m) \end{bmatrix}.$$

Example 3.8 (ℓ_1 norm). The ℓ_1-norm is defined on $\mathbb{R}^n$ as $\|x\|_1 = \sum_{i=1}^n |x_i|$. Using this separability, the proximity operator at $y \in \mathbb{R}^n$ and $\gamma > 0$,

$$\mathbf{prox}_{\gamma\|\cdot\|_1}(y) = \begin{bmatrix} \mathbf{prox}_{\gamma|\cdot|}(y_1) \\ \mathbf{prox}_{\gamma|\cdot|}(y_2) \\ \vdots \\ \mathbf{prox}_{\gamma|\cdot|}(y_m) \end{bmatrix}.$$

For more examples, see (Beck, 2017, Chap. 6) and the website proximity-operator.net.

3.1.4 When no closed form is available

Computing the proximity operator amounts to solving a new problem at each iteration. However, since the problem is made more strongly convex, this new subproblem may be easier to solve. However, this leads to a bi-level implementation. In the following section, we see an intermediate approach that allows not to compute a minimizer of the full function but rather of a piecewise linear approximation.

3.2 Bundle methods

A whole class of method is based on approximating the function using a *bundle* of past information and minimizing this approximation to provide a new point of query for enriching our approximation.

3.2.1 Cutting planes

Essentially, the Subgradient descent uses only once the subgradient information. In other words, the model of F used at iteration k is simply

$$\check{F} : u \mapsto F(x_k) + \langle v_k; u - x_k \rangle.$$

An alternative is to use all the information before k, ie.

$$F_\ell = F(x_\ell) \text{ and } v_\ell \in \partial F(x_\ell) \text{ for } \ell = 0, .., k$$

to form a *cutting plane* model

$$\check{F}_k : u \mapsto \max_{\ell=0,..,k} \left\{ F_\ell + \langle v_\ell; u - x_\ell \rangle \right\}.$$

The function $\check{F}_k$ is

- convex and piecewise linear since this is a maximum of linear functions
- always below F: $\check{F}_k(x) \leq F(x)$ for all x
- increasing with k: $\check{F}_k(x) \leq \check{F}_{k+1}(x)$ for all x

Thus, minimizing $\check{F}_k$ is equivalent to solving the linear problem

$$\begin{aligned} \min_{x,t} \quad & t \\ \text{s.t.} \quad & F_\ell + \langle v_\ell; x - x_\ell \rangle \leq t \text{ for } \ell = 0, .., k \end{aligned}$$

however, this problem may be unbounded, so we need to add a compact convex constraint.

This leads to the cutting planes method

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_{u \in C} \check{F}_k(u) \\ F_{k+1} &= F(x_{k+1}) \\ v_{k+1} &\in \partial F(x_{k+1}) \end{aligned} \qquad \text{(Cutting planes)}$$

where at each iteration a constrained linear cutting plane problem has to be solved.

Theorem 3.9. *Let F be a convex proper lower semi-continuous function and let C be a compact convex set. Assume that F is M Lipschitz and that $\|\partial F(x)\| \leq M$ for all $x \in C$.*

Then, for any $\texttt{tol} > 0$*, after a finite number* k *of iterations*

$$F(x_k) \leq \check{F}_{k-1}(x_k) + \texttt{tol}$$

and then,

$$F(x_k) \leq \min_C F + \texttt{tol}.$$

The theorem thus grants convergence to a minimizer up to some tolerance and also provides a way to stop the algorithm.

Proof. Part 1: We know that $\check{F}_k(x_{k+1}) \leq \min_C F$ since x_{k+1} is the minimizer of the lower model.

Assume for contradiction that $\check{F}_k(x_{k+1}) \leq \min_C F - C$ with $C > 0$ for all k. Since C is compact, we can extract a converging subsequence $(x_{k(\ell)})$ from (x_k). Take ℓ large enough so that $\|x_{k(\ell+1)} - x_{k(\ell)}\| \leq C/(2M)$. Then,

$$\begin{aligned}
\min_C F - C &\geq \check{F}_{k(\ell+1)-1}(x_{k(\ell+1)}) \\
&\geq \check{F}_{k(\ell)}(x_{k(\ell+1)}) \quad \text{(since } \check{F}_k \text{ is non-decreasing)} \\
&\geq F_{k(\ell)} + \langle v_{k(\ell)}; x_{k(\ell+1)} - x_{k(\ell)} \rangle \\
&\geq F_{k(\ell)} - M\|x_{k(\ell+1)} - x_{k(\ell)}\| \\
&\geq \min_C F - C/2
\end{aligned}$$

which is a contradiction thus $\check{F}_k(x_{k+1}) \to \min_C F$.
Part 2: Now, $F(x_k) \geq \min_C F$ by definition.

Similarly, assume for contradiction that $F(x_k) \geq \min_C F + C$ with $C > 0$ for all k and use the same subsequence as above. Then,

$$\begin{aligned}
\min_C F + C &\leq F(x_{k(\ell)}) \\
&= \check{F}_{k(\ell)}(x_{k(\ell)}) \\
&= \check{F}_{k(\ell)}(x_{k(\ell)}) - \check{F}_{k(\ell)}(x_{k(\ell+1)}) + \check{F}_{k(\ell)}(x_{k(\ell+1)}) \\
&\leq M\|x_{k(\ell)} - x_{k(\ell+1)}\| + \check{F}_{k(\ell+1)-1}(x_{k(\ell+1)}) \text{ (since } \check{F}_{k(\ell)} \leq \check{F}_{k(\ell+1)-1}) \\
&\leq C/2 + \min_C F \text{ (since } \check{F}_k(x_{k+1}) \leq \min_C F \text{ for all } k)
\end{aligned}$$

which is again a contradiction thus $F(x_{k(\ell)}) \to \liminf F(x_k) = \min_C F$.
Part 3: Finally, putting the two parts above together, we know that for any $\texttt{tol} > 0$, by taking ℓ sufficiently large, we have

$$\check{F}_{k(\ell)-1}(x_{k(\ell)}) \geq \min_C F - \frac{\texttt{tol}}{2} \quad \text{and} \quad F(x_{k(\ell)}) \leq \min_C F + \frac{\texttt{tol}}{2}$$

and thus

$$F(x_{k(\ell)}) \leq \min_C F + \frac{\texttt{tol}}{2} \leq \check{F}_{k(\ell)-1}(x_{k(\ell)}) + \texttt{tol}$$

which is our stopping criterion. Thus, the algorithm stops in finite time.

Now, we have

$$\begin{aligned} F(x_{k(\ell)}) &\leq \check{F}_{k(\ell)-1}(x_{k(\ell)}) + \texttt{tol} \\ &\leq \check{F}_{k(\ell)-1}(x) + \texttt{tol} \text{ for any } x \in C \text{ (by definition of } x_{k(\ell)}) \\ &\leq F(x) + \texttt{tol} \text{ for any } x \in C \text{ (since } \check{F}_{k(\ell)-1} \text{ is a lower-model)} \end{aligned}$$

which means that $F(x_{k(\ell)}) \leq \min_C F + \texttt{tol}$, which is our result. □

Even though they offer rather good convergence properties and can be very efficient if the function is V-shaped or polyhedral, the cutting planes methods also suffers from numerical instability, increasing computational complexity with the iterations, and can be particularly bad at minimizing some functions.

For instance, consider the function $F(x) = x^2/2$ in one dimension.[15] Then, if $x_0 = 1$, $x_1 = -\varepsilon < 0$, then $x_2 = (1-\varepsilon)/2$. This means that the closest x_1 to the solution, the further x_2. This is a typical *instability* behavior. It is also *not a descent* method since the functional value can increase at each iteration.

[15] This example is taken from (Hiriart-Urruty and Lemaréchal, 1993a, Chap. XV.1.1).

The knowledge of the past values (F_ℓ, x_ℓ, v_ℓ) is called a *bundle* of information and gives its name to the general class of algorithms using them. We will now see how these methods can be stabilized.

3.2.2 Proximal bundle

Now, we give an example of method that uses bundles of information but features:

- a quadratic recall term to take care of the instability behavior;
- a descent test to have a descent method.

The *proximal bundle* method can be written as

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_{u \in C} \check{F}_k(u) + \frac{1}{2\gamma}\|u - \hat{x}_k\|^2 && \text{(Proximal Bundle)} \\ \delta_{k+1} &= F(\hat{x}_k) - \check{F}_k(x_{k+1}) - \frac{1}{2\gamma}\|x_{k+1} - \hat{x}_k\|^2 \\ F_{k+1} &= F(x_{k+1}) \\ v_{k+1} &\in \partial F(x_{k+1}) \\ \hat{x}_{k+1} &= \begin{cases} x_{k+1} & \text{if } F(x_{k+1}) \leq F(\hat{x}_k) - \kappa\delta_{k+1} \quad \text{(serious step)} \\ \hat{x}_k & \text{otherwise} \qquad\qquad\qquad\quad \text{(null step)} \end{cases} \end{aligned}$$

First, notice that since

$$\begin{aligned} -\delta_{k+1} &= \check{F}_k(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - \hat{x}_k\|^2 - F(\hat{x}_k) \\ &\leq \check{F}_k(\hat{x}_k) + \frac{1}{2\gamma}\|\hat{x}_k - \hat{x}_k\|^2 - F(\hat{x}_k) = \check{F}_k(\hat{x}_k) - F(\hat{x}_k) \leq 0, \end{aligned}$$

we indeed have $\delta_{k+1} \geq 0$.

Using the same techniques as before, one can prove that i) there cannot be infinitely many consecutive null steps; and ii) the sequence $(\hat{x}_k)$ minimizes F on C.

Theorem 3.10. *Let F be a convex proper lower semi-continuous function and let C be a closed convex set. Assume that F is M Lipschitz and that $\|\partial F(x)\| \leq M$ for all $x \in C$.*

Then, for any $\gamma > 0$, $\kappa \in (0,1)$, *and* $\texttt{tol} > 0$, *after a finite number* k *of iterations*

$$F(\hat{x}_k) \leq \min_C F + \texttt{tol}.$$

CHAPTER 4 Splitting methods

When a sum of several functions is minimized, some of them may be easier to minimize than others. In addition, optimization objects may be difficult to compute for sums of functions compared to individual ones. In this chapter, we cover how to split optimization objects into atomic ones.

Let us consider problems of the form

$$\min_{x \in \mathbb{R}^n} g(x) + h(x).$$

where g and h are two convex functions, that are potentially nonsmooth.

4.1 The Proximal Gradient

A first simple case is when one of the functions is smooth. In this case, the problem writes as

$$\min_{x \in \mathbb{R}^n} f(x) + g(x). \tag{4.1}$$

where f is smooth and convex, while g is only convex proper and lower semi-continuous.

Since the proximity operator is difficult to compute in general, a rule of thumb is to use a gradient method as soon as possible. Furthermore, in many signal processing or machine learning problems, the objective is of the form $f + g$, with f a smooth loss function that measure the fit between the model and the data and g a nonsmooth regularization, chosen so that the proximity operator is easy to compute.

4.1.1 Algorithm

The *proximal gradient* algorithm consists in iterating

$$x_{k+1} = \mathbf{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right) \qquad \text{(Proximal gradient)}$$

for some $\gamma > 0$ and starting point x_0.

It is worth noticing that this composition can actually be seen as the minimization

of a first-order approximation of f plus g. Indeed:

$$\begin{aligned} x_{k+1} &= \mathbf{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right) \\ &= \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ g(u) + \frac{1}{2\gamma}\|u - x_k + \gamma\nabla f(x_k)\|^2 \right\} \\ &= \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ g(u) + \langle u - x_k, \nabla f(x_k)\rangle + \frac{1}{2\gamma}\|u - x_k\|^2 + \frac{\gamma}{2}\|\nabla f(x_k)\|^2 \right\} \\ &= \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ f(x_k) + \langle u - x_k, \nabla f(x_k)\rangle + g(u) + \frac{1}{2\gamma}\|u - x_k\|^2 \right\} \end{aligned} \tag{4.2}$$

where in the last inequality we remove terms independent of u. We notice that the first two terms approximate f.

This helps us put together the tools for the algorithm's descent lemma.

Lemma 4.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex L-smooth function and let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function. Then, the Proximal gradient method with $\gamma \in (0, 1/L]$ verifies $f(x_{k+1}) + g(x_{k+1}) \leq f(x_k) + g(x_k)$ and*

$$f(x_k) + g(x_k) - (f(x^\star) + g(x^\star)) \leq \frac{\|x^\star - x_0\|^2}{2\gamma k}.$$

Proof. By (4.2), x_{k+1} is the minimizer of the right hand side, which is a $1/\gamma$-strongly convex function, hence for any $z \in \mathbb{R}^n$,

$$\begin{aligned} & f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_k)\rangle + g(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \\ \leq & f(x_k) + \langle z - x_k, \nabla f(x_k)\rangle + g(z) + \frac{1}{2\gamma}\|z - x_k\|^2 - \frac{1}{2\gamma}\|z - x_{k+1}\|^2 \\ \leq & f(z) + g(z) + \frac{1}{2\gamma}\|z - x_k\|^2 - \frac{1}{2\gamma}\|z - x_{k+1}\|^2 \end{aligned}$$

where the second inequality comes from the convexity of f.

Now, the smoothness of f (Lemma 2.2), implies that

$$f(x_{k+1}) \leq f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_k)\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

and using the first set of inequalities, we get

$$\begin{aligned} & f(x_{k+1}) + g(x_{k+1}) \\ & \leq f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_k)\rangle + g(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 + \frac{1}{2}\left(L - \frac{1}{\gamma}\right)\|x_{k+1} - x_k\|^2 \\ & \leq f(z) + g(z) + \frac{1}{2\gamma}\|z - x_k\|^2 - \frac{1}{2\gamma}\|z - x_{k+1}\|^2 + \frac{1}{2}\left(L - \frac{1}{\gamma}\right)\|x_{k+1} - x_k\|^2. \end{aligned}$$

Using $z = x_k$, we get that the sequence of functional values is decreasing and with $z = x^\star$, we obtain the rate with the same proof as for the proximal point method (Theorem 3.4). □

4.1.2 A first splitting method

An alternative (and mode general) construction can be provided for the Proximal gradient, closely linked to *splitting methods*.

First, we notice that a minimizer of problem (4.1) is a point x satisfying

$$0 \in \nabla f(x) + \partial g(x). \tag{4.3}$$

Second, we recall that:

- a gradient step on f leads to $u = x - \gamma \nabla f(x)$
- a proximal step on g generates y such that $y + \gamma \partial g(y) \ni z$

which are our two ingredients.

Now, it suffices to notice that

$$\begin{aligned} & 0 \in \nabla f(x) + \partial g(x) \\ \Leftrightarrow & 0 \in \gamma \nabla f(x) + \gamma \partial g(x) \\ \Leftrightarrow & 0 \in -(x - \gamma \nabla f(x)) + (x + \gamma \partial g(x)) \\ \Leftrightarrow & \begin{cases} u & = x - \gamma \nabla f(x) \\ x + \gamma \partial g(x) & \ni u \end{cases} \end{aligned} \tag{4.4}$$

where we have split the two functions and thus decoupled the difficult problem of finding x verifying (4.3). Thus, we need to iteratively find two points x, u verifying (4.4).

Here, the *order* of the iterates will be important: if we are given x_k, we can compute u_k (by a gradient step), and then a new value x_{k+1} (by a proximal step). We obtain

$$\begin{aligned} & \begin{cases} u_k & = x_k - \gamma \nabla f(x_k) \\ x_{k+1} + \gamma \partial g(x_{k+1}) & \ni u_k \end{cases} \\ \Leftrightarrow & \begin{cases} u_k & = x_k - \gamma \nabla f(x_k) \\ x_{k+1} & = \mathbf{prox}_{\gamma g}(u_k) \end{cases} \\ \Leftrightarrow & x_{k+1} = \mathbf{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)) \end{aligned}$$

which is indeed our Proximal gradient.

4.2 Splitting of two nonsmooth functions

We get back to our original problem

$$\min_{x \in \mathbb{R}^n} g(x) + h(x).$$

where g and h are two convex functions, that are potentially nonsmooth.

A splitting method will find a point x such that $0 \in \partial g(x) + \partial h(x)$ and one can thus wonder when such a point is a minimizer of $g + h$. Fortunately, this is the case under rather mild assumptions.

Lemma 4.2. *Let g and h be two proper lower semi-continuous convex functions. If (i)* $\operatorname{dom} g \cap \operatorname{int} \operatorname{dom} h \neq \emptyset$ *or (ii)* $\operatorname{ri} \operatorname{dom} g \cap \operatorname{ri} \operatorname{dom} h \neq \emptyset$*, then*

$$\partial(g + h) = \partial g + \partial h.$$

Hence, in this section, we will build iterative methods that solve

$$0 \in \partial g(x) + \partial h(x). \tag{4.5}$$

4.2.1 Construction

Since both functions are nonsmooth, our only ingredient is the proximal step:

- a proximal step on g generates x such that $x + \gamma\partial g(x) \ni u$;
- a proximal step on h generates y such that $y + \gamma\partial h(y) \ni z$.

$$
\begin{aligned}
& 0 \in \partial g(x) + \partial h(x) \\
\Leftrightarrow & 0 \in \gamma\partial g(x) + \gamma\partial h(x) \\
\Leftrightarrow & \left\{ \begin{array}{ll} 0 & \in \gamma\partial g(x) + \gamma\partial h(y) \\ x & = y \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{ll} u - x & \in \gamma\partial g(x) \\ y - u & \in \gamma\partial h(y) \\ x & = y \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{ll} u & \in x + \gamma\partial g(x) \\ 2y - u & \in y + \gamma\partial h(y) \\ x & = y \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{ll} x & = \mathbf{prox}_{\gamma g}(u) \\ y & = \mathbf{prox}_{\gamma h}(2y - u) \\ x & = y \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{ll} x & = \mathbf{prox}_{\gamma g}(u) \\ y & = \mathbf{prox}_{\gamma h}(2x - u) \\ x & = y \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{ll} x & = \mathbf{prox}_{\gamma g}(u) \\ y & = \mathbf{prox}_{\gamma h}(2x - u) \\ u & = u + (y - x) \end{array} \right. \qquad (4.6)
\end{aligned}
$$

where the guiding principle was to make the proximity operators appear and then make the iterations computable in order. For the last one, since u is not updated and we have no mean to enforce $x = y$ at that time, we add u on both sides (with an additional parameter > 0).

This way, starting from top right to bottom left, we obtain the following method, called Douglas-Rachford splitting:

$$
\left\{ \begin{array}{ll} x_k & = \mathbf{prox}_{\gamma g}(u_k) \\ y_k & = \mathbf{prox}_{\gamma h}(2x_k - u_k) \\ u_{k+1} & = u_k + (y_k - x_k) \end{array} \right.
$$

which can also be seen as

$$
\begin{aligned}
u_{k+1} &= u_k + (\mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u_k) - u_k) - \mathbf{prox}_{\gamma g}(u_k)) \\
&:= \mathsf{T}(u_k).
\end{aligned}
$$

It is thus interesting to look at the properties of the operator T.

4.2.2 Properties of the operator T

The operator T is a $\mathbb{R}^n \to \mathbb{R}^n$ mapping defined for all $u \in \mathbb{R}^n$ as

$$
\mathsf{T}(u) := u + (\mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma g}(u)).
$$

The goal of this section is to show that T has virtually the same *contraction properties* as a proximity operator. For this recall from Proposition 3.2 that for any convex lower semi-continuous proper function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, then the following propositions are equivalent:

i) $x = \mathbf{prox}_{\gamma g}(y)$;
ii) $(y - x)/\gamma \in \partial g(x)$;
iii) $g(u) \geq g(x) + \langle (y - x)/\gamma, u - x \rangle$ for any $u \in \mathbb{R}^n$.

This enables us to show the following lemma, which highlights the proximity operator is *firmly non-expansive*.

Lemma 4.3. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then for any $u, z \in \mathbb{R}^n$, $\gamma > 0$,*

$$\begin{aligned} &\|\mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z)\|^2 \leq \langle u - z, \mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z) \rangle \\ \Leftrightarrow &\|\mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z)\|^2 \leq \|u - z\|^2 - \|u - \mathbf{prox}_{\gamma g}(u) - z + \mathbf{prox}_{\gamma g}(z)\|^2 \end{aligned}$$

Proof. Take point (iii) above with $y \leftarrow u, u \leftarrow \mathbf{prox}_{\gamma g}(z)$ and $y \leftarrow z, u \leftarrow \mathbf{prox}_{\gamma g}(u)$. Summing both inequalities gives the first result. The second one comes directly afterwards. □

Now, we can prove the same result for the Douglas-Rachford operator T.

Theorem 4.4. *Let $g, h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be two convex lower semi-continuous proper functions, then for any $u, z \in \mathbb{R}^n$, $\gamma > 0$,*

$$\begin{aligned} &\|\mathrm{T}(u) - \mathrm{T}(z)\|^2 \leq \langle u - z, \mathrm{T}(u) - \mathrm{T}(z) \rangle \\ \Leftrightarrow &\|\mathrm{T}(u) - \mathrm{T}(z)\|^2 \leq \|u - z\|^2 - \|u - \mathrm{T}(u) - z + \mathrm{T}(z)\|^2 \end{aligned}$$

Proof. Let $u, z \in \mathbb{R}^n$,

$$
\begin{aligned}
&\langle \mathsf{T}(u) - \mathsf{T}(z), u - z \rangle \\
&= \langle u - z, u - z \rangle \\
&\quad + \langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), u - z \rangle \\
&\quad - \langle \mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z), u - z \rangle \\
&\geq \|u - z\|^2 - 2\langle \mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z), u - z \rangle + \|\mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z)\|^2 \\
&\quad + \langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), u - z \rangle \\
&= \|u - \mathbf{prox}_{\gamma g}(u) - (z - \mathbf{prox}_{\gamma g}(z))\|^2 \\
&\quad + \langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), u - z \rangle \\
&= \|u - \mathbf{prox}_{\gamma g}(u) - (z - \mathbf{prox}_{\gamma g}(z))\|^2 \\
&\quad - \langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), 2\mathbf{prox}_{\gamma g}(u) - u - (2\mathbf{prox}_{\gamma g}(z) - z) \rangle \\
&\quad + \langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), 2\mathbf{prox}_{\gamma g}(u) - 2\mathbf{prox}_{\gamma g}(z) \rangle \\
&\geq \|u - \mathbf{prox}_{\gamma g}(u) - (z - \mathbf{prox}_{\gamma g}(z))\|^2 \\
&\quad - 2\langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z), 2\mathbf{prox}_{\gamma g}(u) - u - (2\mathbf{prox}_{\gamma g}(z) - z) \rangle \\
&\quad + \|\mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z)\|^2 \\
&\quad + 2\langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(\mathbf{prox}_{\gamma g}(z) - z), \mathbf{prox}_{\gamma g}(u) - \mathbf{prox}_{\gamma g}(z) \rangle \\
&= \|u - \mathbf{prox}_{\gamma g}(u) - (z - \mathbf{prox}_{\gamma g}(z))\|^2 \\
&\quad + 2\langle \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(\mathbf{prox}_{\gamma g}(z) - z), u - \mathbf{prox}_{\gamma g}(u) - (z - \mathbf{prox}_{\gamma g}(z)) \rangle \\
&\quad + \|\mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - z)\|^2 \\
&= \|u + \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(u) - u) - \mathbf{prox}_{\gamma g}(u) - (z + \mathbf{prox}_{\gamma h}(2\mathbf{prox}_{\gamma g}(z) - \mathbf{prox}_{\gamma g}(z))\|^2 \\
&= \|\mathsf{T}(u) - \mathsf{T}(z)\|^2
\end{aligned}
$$

where the inequalities come from Lemma 4.3. □

4.2.3 Convergence

Now that we have highlighted our base property, let us show the convergence of our method. Note that this kind of proof is ubiquitous in optimization with *monotone operators*.

Theorem 4.5. *Let $g, h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be two convex lower semi-continuous proper functions such that:*

- *(i)* $\operatorname{dom} g \cap \operatorname{int} \operatorname{dom} h \neq \emptyset$ *or (ii)* $\operatorname{ri} \operatorname{dom} g \cap \operatorname{ri} \operatorname{dom} h \neq \emptyset$ *;*
- *$g + h$ has a minimizer.*

Then, the Douglas-Rachford algorithm

$$
\left\{
\begin{array}{rl}
x_k &= \mathbf{prox}_{\gamma g}(u_k) \\
y_k &= \mathbf{prox}_{\gamma h}(2x_k - u_k) \\
u_{k+1} &= u_k + (y_k - x_k)
\end{array}
\right.
$$

produces a sequence (x_k) that converges to a minimizer of $g + h$.

Proof. The two assumptions in the theorem imply the existence of minimizers that are the points verifying $0 \in \partial g(x) + \partial h(x)$ (see Eq. (4.5)).

Let $x^\star$ be a point such that $0 \in \partial g(x^\star) + \partial h(x^\star)$. Then, by Eq. (4.6), there is a $u^\star$ such that:

- $u^\star = \mathsf{T}(u^\star)$, i.e., $u^\star$ is a fixed point of T;
- $x^\star = \mathbf{prox}_{\gamma g}(u^\star)$.

Then, from Theorem 4.4, we get that

$$\|u_{k+1} - u^\star\|^2 = \|\mathsf{T}(u_k) - \mathsf{T}(u^\star)\|^2 \leq \|u_k - u^\star\|^2 - \|\mathsf{T}(u_k) - u_k\|^2$$

which means that $(\|u_k - u^\star\|^2)$ is non-increasing and thus convergent.

Also, since $\|u_k - u^\star\|^2 \leq \|u_0 - u^\star\|^2$, (u_k) is bounded and thus has a converging subsequence. Furthermore, the limit u of this subsequence must verify $\mathsf{T}(u) - u = 0$ (i.e., it has to be a fixed point of T).

Since u is a fixed point of T, we can replace $u^\star$ above by u and obtain that $(\|u_k - u\|^2)$ converges, to 0.

Finally, we get that (u_k) converges to u. This implies that (x_k) converges to $x = \mathbf{prox}_{\gamma g}(u)$, which is a point verifying $0 \in \partial g(x) + \partial h(x)$ by Eq. (4.6). □

Bibliography

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1993a. Two volumes.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1993b. Two volumes.

Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, 1998.

Kôsaku Yosida. *Functional analysis*, volume 123. springer, 1988.