Numerical Optimization
Part I

Franck Iutzeler

January 28, 2022

Contents

CHAPTER 1 VARIATIONAL ANALYSIS

The purpose of this first part is to properly introduce the notations and the notions of differentiability at play when dealing with functions. We will thus see how to define subgradients, and gradients when the function is differentiable. Finally, we will take a first look at the gradient algorithm.

In the first page of the renowned book "Variational analysis" by R. Tyrrell Rockafellar and Roger J-B Wets (Rockafellar and Wets, 1998), we are told that "it's convenient for many purposes to consider functions F that are allowed to be extended-real-valued, i.e., to take values in $\overline{\mathbb{R}} = [-\infty, +\infty]$ instead of just $\mathbb{R} = (-\infty, +\infty)$", we will thus adopt this convention ourselves.

A fundamental question in variational analysis is the study of the minimum (or equivalently maximum) of functions defined over a Euclidean space $\mathbb{R}^n$. In all this course, we will place ourselves in the (finite-dimensional) Euclidean space $\mathbb{R}^n$, with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $x \mapsto \|x\| := \sqrt{\langle x, x \rangle}$.

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.

1.1 SUBGRADIENTS

In order to investigate the local behavior of a function with respect to minimization, a first natural step is to consider local affine lower approximations. This *first-order* information is captured by the notion of subgradients. There is a variety of subgradients

and several ways to express them, see (Rockafellar and Wets, 1998, Chap. 7,8), (Mordukhovich, 2006, Chap. 1) for general references. We give here only the notions that will be used for our purposes following the terminology and notations of (Rockafellar and Wets, 1998, Chap. 8).

Definition 1.1 (Subgradients). Consider a function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which $F(x)$ is finite:

- the set of *regular subgradients* is defined as

$$\widehat{\partial} F(x) = \{v : F(u) \geq F(x) + \langle v, u - x\rangle + o(\|u - x\|) \text{ for all } u \in \mathbb{R}^n\}. \tag{1.1}$$

- the set of *(general or limiting) subgradients* is defined as

$$\partial F(x) = \left\{\lim_r v_r : v_r \in \widehat{\partial} F(u_r),\ u_r \to x,\ F(u_r) \to F(x)\right\}. \tag{1.2}$$

If $F(x)$ is infinite, $\widehat{\partial} F(x) = \partial F(x) = \emptyset$.

While the regular subgradient seems simpler and more appealing at first, we will use the general subgradient in all the following, simply referenced under the name subgradient for simplicity. The reason for this is its superior continuity properties as stated in the following lemma.

Lemma 1.2 (Rockafellar and Wets 2009, Th. 8.6, Prop. 8.7 [⋆]). *Consider a function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which $F(x)$ is finite, then the sets of regular subgradients $\widehat{\partial} F(x)$ and general subgradients $\partial F(x)$ are closed. Furthermore, the set of general subgradients ∂F is outer semi-continuous at x, ie.*

$$\limsup_{u \to x \text{ with } F(u) \to F(x)} \partial F(u) := \{v : \exists u_r \to x, \exists\, v_r \to v \text{ with } v_r \in \partial F(u_r)\} \subset \partial F(x)$$

Note that the regular and limiting subdifferentials at some point x coincide in a variety of situations, we then say that the function is *(Clarke) regular* at x (Rockafellar and Wets, 2009, Def. 7.25, Cor. 8.11). While less natural in its definition, the outer semi-continuity property of the general subgradient allows us, for example, to deduce that any limit point x of a sequence (x_k) satisfy $0 \in \partial F(x)$ if the distance from $\partial F(x_k)$ to 0 vanishes.

The condition $0 \in \partial F(x)$ is particularly interesting since it is related to local minimas by Fermat's rule.

Theorem 1.3 (Fermat's rule). *If a proper function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ has a local minimum at x (ie. if there is a neighborhood $\mathcal{U}$ of x such that $F(x) \leq F(u)$ for all $u \in \mathcal{U}$) then $0 \in \partial F(x)$.*

1.2 Differentiability

Differentiability plays a central role in optimization. This is somehow a special case of the notion of subgradient defined above but the treatment of differentiable functions will be rather different algorithmically. In order to promote even more this difference, we will adopt the following convention for the name of generic functions: (*i*) f if it is differentiable; (*ii*) g if it is not assumed differentiable; and (*iii*) F if the differentiability does not play a role in the result.

1.2.1 Derivative of a function from $\mathbb{R}$ to $\mathbb{R}$

In this basic case, the notion of differentiability is quite direct.

Definition 1.4. A function $f : \mathcal{V} \subset \mathbb{R} \to \mathbb{R}$ defined on a open subset[1] $\mathcal{V}$ of $\mathbb{R}$ is differentiable at $x \in \mathcal{V}$ if the derivative (ie. the limit)

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

exists. This function f is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$.

[1] At first read, you can take $\mathcal{V}$ as the full space to fix ideas

This definition is equivalent to the existence of a real number $f'(x)$ such that

$$f(x+h) = f(x) + f'(x)h + o(|h|).$$

Note that we now only consider an open subset of $\mathbb{R}$ over which the function is finite-valued. If f takes infinite values on any open set containing x, then it cannot be differentiable at that point.

In addition, if f is differentiable at x, it is necessarily continuous at x. The derivative f' is itself a function from $\mathcal{V} \to \mathbb{R}$ and may also be continuous (on $\mathcal{V}$), in which case, we say that f is continuously differentiable, often denoted $C^1(\mathcal{V})$ or simply C^1.

The derivative of the derivative is called the second-order derivative, noted f''. If it exists and is continuous, we say that f is C^2. Iterating, we can easily define higher order derivatives and differentiability classes up to C^∞.

1.2.2 Gradient of a function from $\mathbb{R}^n$ to $\mathbb{R}$

Let us now consider a function defined over an open subset $\mathcal{V}$ of $\mathbb{R}^n$

$$\begin{array}{rccc} f: & \mathcal{V} \subset \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x = [x_1, .., x_n] & \longmapsto & f(x) \end{array}.$$

For every $x \in \mathcal{V}$, the i-th *partial function* is defined on $\mathcal{V}' \subset \mathbb{R}$ as

$$\begin{array}{rccc} \phi_{i,x}: & \mathcal{V}' & \longrightarrow & \mathbb{R} \\ & u & \longmapsto & f(x_1, .., x_{i-1}, u, x_{i+1}, .., x_n) \end{array},$$

and since this function falls into the case of the previous section, we can study its differentiability. If for all i, $\phi_{i,x}$ is differentiable at x_i, then, we will say that f is differentiable at x.

Definition 1.5. A function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^n$ is differentiable at $x \in \mathcal{V}$ if for all $i = 1, .., n$, the derivative (ie. the limit)

$$\frac{\partial f}{\partial x_i}(x) := \lim_{h \to 0} \frac{\phi_{i,x}(x_i + h) - \phi_{i,x}(x_i)}{h}$$

exists. This function f is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$. Further, if f is differentiable on $\mathcal{V}$, we define its *gradient* as the $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}^n$ mapping

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

Similar to what was obtained in the one-dimensional case, we have a *first-order* development of f at a point x at which f is differentiable:

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

1.2.3 Jacobian of a mapping $\mathbb{R}^m$ to $\mathbb{R}^n$

Now, let us consider the case of a mapping (ie. a multi-valued function) from $\mathbb{R}^m$ to $\mathbb{R}^n$

$$\begin{array}{rccc} c: & \mathcal{V} \subset \mathbb{R}^m & \longrightarrow & \mathbb{R}^n \\ & x = [x_1, .., x_m] & \longmapsto & c(x) = [c_1(x), .., c_n(x)] \end{array}.$$

A mapping is differentiable if and only if each of its *component functions* is differentiable as formalized in the following definition.

Definition 1.6. A mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^m$ is differentiable at $x \in \mathcal{V}$ if for all $i = 1, .., n$, and all $j \in 1, .., m$, the derivative $\frac{\partial c_i}{\partial x_j}(x)$ exists. This mapping c is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$. Further, if c is differentiable on $\mathcal{V}$, we define its *Jacobian* as the $\mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ mapping[2]

[2] The name comes from Carl Gustav Jacob Jacobi (1804-1851), a German mathematician.

$$Jc(x) = \begin{bmatrix} \nabla c_1(x)^\top \\ \vdots \\ \nabla c_n(x)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial c_1}{\partial x_1}(x) & \dots & \frac{\partial c_1}{\partial x_m}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial c_n}{\partial x_1}(x) & \dots & \frac{\partial c_n}{\partial x_m}(x) \end{bmatrix}.$$

While, we do not often differentiate mappings, we often differentiate compositions of a function and mapping. For this, the *chain rule* gives a efficient formula based on the respective gradients and Jacobian of the functions.

Lemma 1.7 (Chain rule)**.** *Take a function $f : \mathcal{V}' \subset \mathbb{R}^n \to \mathbb{R}$ and a mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$. If c is differentiable at $x \in \mathcal{V}$ and f is differentiable at $c(x) \in \mathcal{V}'$, then $f \circ c$ is differentiable at x and its gradient can be obtained by*[3]

[3] $f \circ c(x) = f(c(x))$

$$\nabla f \circ c(x) = Jc(x)^\top \nabla f(c(x)). \qquad \text{(Chain rule)}$$

The first-order development of $f \circ c$ is thus

$$f \circ c(x+h) = f \circ c(x) + \langle Jc(x)^\top \nabla f(c(x)), h \rangle + o(\|h\|).$$

1.2.4 Second-order differentiability

The derivative of the gradient, that is the second-order derivative of the function, is often used in numerical optimization methods.

Definition 1.8. A function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}$ is twice differentiable at $x \in \mathcal{V}$ if its gradient is differentiable at $x \in \mathcal{V}$.
Further, if f is twice differentiable on $\mathcal{V}$, we define its *Hessian* as the $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ mapping[4]

[4] also denoted by Hf, its name comes from Ludwig Otto Hesse (1811-1874), a German mathematician.

$$\nabla^2 f(x) = J\nabla f(x) = \begin{bmatrix} \frac{\partial^2 f}{(\partial x_1)^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \dots & \frac{\partial^2 f}{(\partial x_n)^2}(x) \end{bmatrix}.$$

This definition comes with the following important property.

Lemma 1.9. *The Hessian of a function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathcal{V}$ is a symmetric matrix.*

Proof. This follows directly from Schwarz's theorem.[5] □

[5] Hermann Schwarz (1843-1921), German mathematician, was the first to propose a rigorous proof of the symmetry of second derivatives (also called the equality of mixed partials).

Remark 1.10 (Hessian at a local minimum). If f admits a local minimum at x and is twice differentiable at x, then $\nabla f = 0$ by Fermat's rule (Theorem 1.3) but we can also show that $\nabla^2 f(x)$ is positive semi-definite; see Tutorial 1 Exercise 1.3. ◀

1.2.5 Fréchet derivatives [⋆]

The notion of Fréchet derivatives generalizes the notion of gradient and Jacobian seen above. A mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^m$ is *Fréchet differentiable* at $x \in \mathcal{V}$ if there exists a linear operator

$$\begin{array}{rccc} \mathrm{D}c(x) : & \mathbb{R}^m & \longrightarrow & \mathbb{R}^n \\ & h & \longmapsto & \mathrm{D}c(x)[h] \end{array}$$

called the (Fréchet) *differential* of c at x,[6] such that

[6] from Maurice René Fréchet (1878-1973), a French mathematician.

$$c(x+h) = c(x) + \mathrm{D}c(x)[h] + o(\|h\|)$$
$$\text{or, equivalently } \lim_{h \to 0} \frac{\|c(x+h) - c(x) - \mathrm{D}c(x)[h]\|}{\|h\|} = 0.$$

Then, if f is a $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ function, the gradient of f can be defined as the unique element of $\mathbb{R}^n$ that satisfies

$$\mathrm{D}f(x)[h] = \langle \nabla f(x), h \rangle \text{ for all } h \in \mathbb{R}^n$$

and thus, in line with the regular subgradient notation, it can also be defined as

$$\nabla f(x) = \{v : f(u) = f(x) + \langle v, u - x \rangle + o(\|u - x\|) \text{ for all } u \in \mathbb{R}^n\}. \tag{1.3}$$

The same can be done for mappings and the Jacobian of c can be defined as the unique $\mathbb{R}^n \times \mathbb{R}^m$ operator $Jc(x)$ such that $\mathrm{D}c(x)[h] = Jc(x)h$.

Finally, the Chain rule for differentials is

$$\mathrm{D}(f \circ c)(x)[h] = \mathrm{D}f(c(x))[\mathrm{D}c(x)[h]] = \langle \nabla f(c(x)), Jc(x)h \rangle = \langle Jc(x)^\top \nabla f(c(x)), h \rangle.$$

1.2.6 Link with subdifferentials

To be complete, let us relate the notion of gradient defined here with the subdifferentials defined before.

Lemma 1.11. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which f is differentiable, then $\nabla f(x) = \widehat{\partial} f(x) \subset \partial f(x)$. If, in addition, f is continuously differentiable around x, then $\nabla f(x) = \partial f(x)$.*

Proof. For the first part, interpret directly (1.3) as (1.1). For the second part, the continuity of ∇f enables leaves no other choice for a limit in (1.2) than $\nabla f(x)$. □

In the common case, where we deal with the sum of two functions, the following lemma is particularly useful.

Lemma 1.12. *If $F = f + g$ with f continuously differentiable around x and $g(x)$ finite, then $\partial F(x) = \nabla f(x) + \partial g(x)$.*

Proof. Direct from the definitions. □

1.3 Smoothness and Gradient descent

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 1.13. 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

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

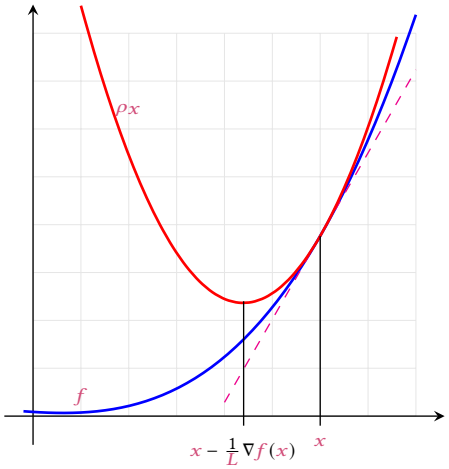
From this property, we can derive this highly important lemma.

Lemma 1.14. *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*

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

Proof. See Tutorial 1. □

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 1.14 gives us

$$f(u) \leq f(x) - \left(\frac{1}{\gamma} - \frac{L}{2}\right) \|x - u\|^2 = f(x) - \left(\gamma - \frac{L\gamma^2}{2}\right) \|\nabla f(x)\|^2. \tag{1.4}$$

Thus, taking a gradient step leads to a strict functional decrease ($f(u) < f(x)$) as soon as $\gamma < 2/L$. This is the core idea behind the *gradient descent* algorithm.[7] Take $x_0 \in \mathbb{R}^n$ and $\gamma > 0$, the gradient descent algorithm consists in iterating

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

and leads to the following guarantees.

[7] introduced by Louis Augustin Cauchy (1789–1857), a French mathematician, in his "Compte Rendu à l'Académie des Sciences" of October 18, 1847.

Theorem 1.15. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with a L-Lipschitz continuous gradient and such that* $\inf f > -\infty$. *Assume that* (Gradient descent) *is run with $0 < \gamma < 2/L$, then $(f(x_k))$ converges and any limit point $\overline{x}$ of (x_k) satisfies $\nabla f(\overline{x}) = 0$.*

Proof. See Tutorial 1. □

Even though the above theorem is only a partial justification, gradient descent is widely used for finding critical points of smooth functions. The link between finding critical points and minimizing a function will be brought in the next chapter by convexity. In that case, the guarantees of gradient descent will be strengthened.

Finally, let us conclude this chapter with a quote from the original paper by Cauchy (Cauchy et al., 1847) that also applies to us *"I'll restrict myself here to outlining the principles underlying [my method], with the intention to come again over the same subject"*.[8]

[8] In the original text: *" Je me bornerai pour l'instant à indiquer les principes sur lesquels [ma méthode] se fonde, me proposant de revenir avec plus de détails sur le même sujet, dans un prochain mémoire."*. The translation and reference is due to Claude Lemaréchal, see (Lemaréchal, 2012).

CHAPTER 2 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*.

2.1 CONVEX SETS

2.1.1 Motivation: Projecting onto a closed set

Similarly to orthogonal projections onto affine subspaces, we can define projection on *nonempty closed* sets.[9]

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

$$\inf_{x\in C} F_y(x) := \frac{1}{2}\|y-x\|^2 \tag{2.1}$$

which intuitively amounts to projecting y onto C.

First, take $u \in C$, and define $S := \{x \in \mathbb{R}^n : \|y-x\|^2 \leq \|y-u\|^2\}$. Then, the problem (2.1) is equivalent to

$$\inf_{x\in C\cap S} F_y(x) := \frac{1}{2}\|y-x\|^2 \tag{2.2}$$

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 2.1. *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 $\bar{\beta} := \inf_{x\in S} F(x) > -\infty$. 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 (2.2), and thus of (2.1), 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.[10]

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

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

2.1.2 Convexity for sets

Let us now introduce the definition of a convex set.

Definition 2.2. 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 $\{x : \|x - s\| \le r\}$
- Half spaces $\{x : \langle s, x\rangle \le r\}$ and open half spaces $\{x : \langle s, x\rangle < r\}$
- Simplices $\{x : \sum_{i=1}^n x_i = 1 \text{ and } x_i \ge 0 \text{ for all } i = 1, .., n\}$
- 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"

2.1.3 Projection on convex sets

Getting back to the projection problem (2.1)

$$\min_{x \in C} F_y(x) := \frac{1}{2}\|y - x\|^2 \tag{2.3}$$

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

Suppose that $x_1^\star \neq x_2^\star$ are two distinct solutions of (2.3). 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 2.3. *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 (2.3).*

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

Theorem 2.4. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, $\text{proj}_C(y)$ is the projection of y onto C if and only if*

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

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

2.1.4 Minimization over convex sets

Now, let us consider a more general problem: minimizing a 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

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

We directly note that if C is empty, the problem is impossible[11] 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.

[11] *infeasible* in the optimization language.

The *constrained* variant of Fermat's rule (Theorem 1.3) that links the (sub)gradient of the function with local minimas writes as follows.

Theorem 2.5 ((Rockafellar and Wets, 1998, Th. 6.12,8.15))**.** *If a proper lower-semicontinuous function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ has a local minimum at x constrained to the convex set C (ie. if there is a neighborhood $\mathcal{U}$ of x in C such that $F(x) \leq F(u)$ for all $u \in \mathcal{U}$) then $0 \in \partial F(x) + N_C(x)$ or,[12] equivalently,*

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

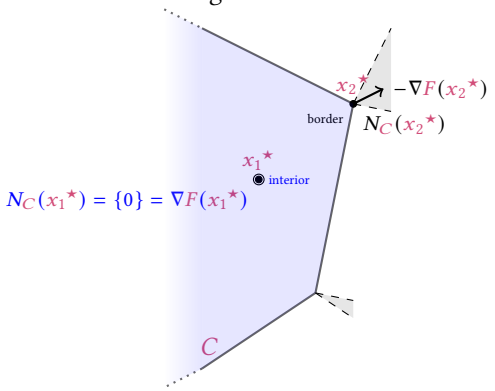
for any $v \in \partial F(x)$ and all $y \in C$.
In particular, if F is differentiable, $0 \in \nabla F(x) + N_C(x)$ means that

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

for all $y \in C$.

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

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



2.2 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.

2.2.1 Definition

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

A function is convex if and only if its *epigraph*[13] is convex. However, the following definition is much more direct.

Definition 2.6. A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if for any $x, u \in \operatorname{dom} F$, $F((1-\alpha)x + \alpha u) \leq (1-\alpha)F(x) + \alpha F(u)$ 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 (for some, we will prove that they preserve convexity in Tutorial 2):

- 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 2.7. *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. See Tutorial 2. □

2.2.2 Subgradients 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 again captured by the notion of subgradients.

Lemma 2.8 (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,*

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

Thus, F is regular at any point 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.

Furthermore, we have the same link between subgradients and optimality when constrained to a convex set.

Theorem 2.9 ((Rockafellar and Wets, 1998, Th. 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 $0 \in \partial F(x) + N_C(x)$ or, equivalently,*

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

for any $v \in \partial F(x)$ and all $y \in C$.

2.2.3 Differentiable convex functions

First, Theorem 2.9 can be a little simplified if the function is differentiable.

Theorem 2.10 ((Rockafellar and Wets, 1998, Th. 6.12)). *Consider a proper lower-semicontinuous convex and differentiable 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 $0 \in \nabla f(x) + N_C(x)$ which means that*

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

for all $y \in C$.

In addition, for a differentiable f, convexity can be seen directly as a property on the gradient mapping.

Theorem 2.11 (Bauschke and Combettes 2011, Prop. 17.10). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function with open domain.[14] 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.

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

2.2.4 Strict 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 2.12. *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 $\mathbb{R}^n$.*

Proof. See Tutorial 2. □

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.

2.2.5 Strong convexity

While convexity provides affine lower bounds, having quadratic lower-bounds enable to get a better control that may have a great impact on the convergence of optimization methods; this is captured by the notion of strong convexity.

Definition 2.13. 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 $\tilde{F} := F - \frac{1}{2}\mu\|\cdot\|^2$ is convex and verifies $\partial\tilde{F} = \partial F - \mu\cdot$ by Lemma 1.12, we get that for any $x \in \mathbb{R}^n$ and any $v \in \partial F(x)$

$$F(u) \geq F(x) + \langle v, u - x\rangle + \frac{\mu}{2}\|u - x\|^2 \text{ for all } u \in \mathbb{R}^n \tag{2.4}$$

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 2.14. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strongly convex lower semi-continuous proper function and C a convex set, then F has exactly one minimizer on C. In particular, F has exactly one minimizer 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 (2.4), 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)). This means that 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 2.1 to this compact set gives us the existence of a minimal value, which is unique from the quadratic lower bound expressed in (2.4). □

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

Theorem 2.15. *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.

2.3 Back to the gradient algorithm

We saw in Section 1.3 that the (Gradient descent) algorithm on a L-smooth 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 $\gamma \in (0, 2/L)$.

In Chapter 1, we saw that the functional values were decreasing and all limit points where critical points of f. However, we had no convergence guarantee and no rate. Convexity will help us get these rates. For this part, our main reference will be (Bubeck et al., 2015, Chap. 3.2,3.4).

2.3.1 Gradient algorithm for convex functions

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

Theorem 2.16. *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;*[15]
- *(rate) $f(x_k) - f(x^\star) \leq \dfrac{2L\|x_0 - x^\star\|^2}{k}$ for any minimizer $x^\star$ of f.*

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

Proof. See Tutorial 4. □

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 2.17 (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 $\alpha_{k+1} = (k+2)/(k+3)$.[16]

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

2.3.2 Gradient algorithm for strongly convex functions

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

Theorem 2.18. *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;*[17]
- *(rate) $f(x_k) - f(x^\star) \leq \left(\dfrac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x_0 - x^\star\|^2$ where $\kappa = \frac{L}{\mu} \geq 1$.*

[17] *unique by strong convexity*

Proof. See Tutorial 4. □

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).

2.3.3 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 2.19. *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;*[18]
- *(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.*

[18] *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 $y \in C$.*

Proof. We use Theorem 2.4 to get that since $x_{k+1} = \text{proj}_C\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$, we have

$$\langle x_k - \frac{1}{L}\nabla f(x_k) - x_{k+1}, z - x_{k+1}\rangle \geq 0 \text{ for any } z \in C$$

and taking $z = x_k$ this gives

$$\langle \nabla f(x_k), x_k - x_{k+1}\rangle \leq \underbrace{\langle L(x_k - x_{k+1}), x_k - x_{k+1}\rangle}_{:= g_C(x_k)}.$$

Then, smoothness of f implies that

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

and the rest of the proof is similar to the one of Theorem 2.16 with $\nabla f(x_k)$ replaced with $g_C(x_k)$. □

CHAPTER 3 Nonsmooth minimization and the Proximity operator

The gradient method is very helpful for minimizing smooth functions. However, when a function is non-differentiable, the (sub)gradient method has very degraded performances. Fortunately, in many applications of interest the nonsmooth functions have some particular structure that enable us to use a powerful tool: the proximity operator.

In this chapter, we consider the problem

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

where g is a nonsmooth convex function.

3.1 The Subgradient method

A direct method to minimize a convex nonsmooth function is to mimic the gradient method and to subgradient descent:

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

However, these iterations may not converge. For instance, take $g = |\cdot|$, then (x_k) will oscillate around 0 for any $\gamma > 0$.

In fact, we have the following result.

Theorem 3.1. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function with $\|v\| \leq M$ for any $x \in \operatorname{dom} g$ and any $v \in \partial g(x)$. Then, (Subgradient descent) started with x_0 such that $\|x_0 - x^\star\| \leq R$ with $\gamma = \frac{R}{M\sqrt{k}}$ generates iterates that verify*

$$g\left(\frac{1}{k}\sum_{t=0}^{k-1} x_t\right) - g(x^\star) \leq \frac{MR}{\sqrt{k}}.$$

Proof. See (Bubeck et al., 2015, Th. 3.2). □

This is not completely satisfying since the stepsize is decreasing which limits the rate and the iterates do not converge. Thankfully, we have a better tool to deal with certain nonsmooth functions.

3.2 The Proximity Operator

3.2.1 Definition

A central tool to tackle non-differentiable functions is the *proximity operator*, introduced by (Moreau, 1965), and denoted $\mathbf{prox}_{\gamma g}$ for a step-size $\gamma > 0$ and a nonsmooth function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$; it is defined as the set-valued mapping

$$\mathbf{prox}_{\gamma g}(y) := \operatorname{argmin}_{u \in \mathbb{R}^n} \Big\{ \underbrace{g(u) + \frac{1}{2\gamma}\|u - y\|^2}_{:=\rho_y(u)} \Big\}.$$

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

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

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

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

which mirrors (1.4) (the descent inequality of a gradient step on a smooth function) but for a nonsmooth function.[19]

With this respects, the proximity operator provides a alternative to the use of subgradients since they are not able to provide descent inequalities such as (1.4) 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.

[19] 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 $e_\gamma g(y) = \inf_{u \in \mathbb{R}^n} \rho_y(u)$ (Moreau, 1965; Yosida, 1988).

3.2.2 Properties

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

Theorem 3.2. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $\mathbf{prox}_{\gamma g}(y)$ is a singleton for any $\gamma > 0$ and any $y \in \mathbb{R}^n$.*

Proof. Since g is convex, ρ_y is $\frac{1}{\gamma}$-strongly convex. Then, Lemma 2.14 guarantees the existence and uniqueness of the minimizers of $\rho_y(u)$ for any u, which means that $\mathbf{prox}_{\gamma g}(y)$ is well-defined and unique. □

In addition, we have the following identity.

Proposition 3.3. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then the following propositions are equivalent:*

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

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

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

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.

Corollary 3.4. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then for any $y, z \in \mathbb{R}^n$*

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

Proof. See Tutorial 3. □

3.2.3 Convergence of the proximal point algorithm

Now, let us investigate the proximal point algorithm:

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

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

Corollary 3.5. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $x^\star$ is a minimizer of g if and only if $x^\star = \mathbf{prox}_{\gamma g}(x^\star)$ (for any $\gamma > 0$).*

Proof. From Proposition 3.3, we have that $x^\star = \mathbf{prox}_{\gamma g}(x^\star)$ if and only if $0 \in \partial g(x^\star)$ which is equivalent to $x^\star$ being a minimizer of g since it is convex. □

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

Theorem 3.6. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function. Then, the (Proximal Point) method with $\gamma > 0$ verifies $g(x_{k+1}) \leq g(x_k)$ and*

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

Proof. First, since $x_{k+1} = \mathbf{prox}_{\gamma g}(x_k)$,

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

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

Since $x_{k+1} = \mathbf{prox}_{\gamma g}(x_k)$, it is the minimum of the $1/\gamma$-strongly convex function ρ_{x_k},[20] thus

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

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

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

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

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

which gives the result. □

[20] If $x^\star$ is the minimizer of a μ-strongly convex function F, then $0 \in \partial F(x^\star)$ and (2.4) gives us that $F(x^\star) \leq F(u) - \frac{\mu}{2}\|u - x^\star\|^2$.

3.2.4 Examples

Example 3.7 (Squared norm). For $g(x) = \frac{1}{2}\|x\|^2$, the proximity operator can be computed explicitly. Since $\rho_y : u \mapsto g(u) + \frac{1}{2\gamma}\|u - y\|^2$ is strongly convex, there is a unique minimizer x and it verifies $\nabla s(x) = 0$. 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 3.8 (Absolute value). The proximity operator of the absolute value admits a closed form expression: for $y \in \mathbb{R}$ and $\gamma > 0$,

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

See Tutorial 3.

[21] By coordinates, or blocks of coordinates.

A very useful calculus rule for the proximity operator is that if g is separable:[21]

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

then the proximity operator of g can be obtained from those of the (g_i):

$$\mathbf{prox}_{\gamma g}(y_1, y_2, .., y_m) = \begin{bmatrix} \mathbf{prox}_{\gamma g_1}(y_1) \\ \mathbf{prox}_{\gamma g_2}(y_2) \\ \vdots \\ \mathbf{prox}_{\gamma g_m}(y_m) \end{bmatrix}.$$

Example 3.9 (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.

3.2.5 Relation with the Projection operator

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[22]

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

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

Lemma 3.10. *Let $C \subset \mathbb{R}^n$ be a non-empty closed convex set. Then, for any $\gamma > 0$,*

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

Proof. It is enough to remark that the inner minimization in $\mathbb{R}^n$ amounts to minimizing over C since otherwise the inner value is $+\infty$. This exactly gives the definition of the projection. □

3.3 The Proximal Gradient

Now that we have a tool to handle simple nonsmooth functions, we can address the minimization of the sum of a smooth function and a nonsmooth one:

$$\min_{x \in \mathbb{R}^n} f(x) + g(x). \tag{3.2}$$

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.

3.3.1 Motivation: Splitting [⋆]

In the convex case, a minimizer of problem (3.2) is a point x satisfying

$$0 \in \nabla f(x) + \partial g(x). \tag{3.3}$$

In order to decouple these two functions, *splitting* methods have been developed. They consist in finding a point satisfying (3.3) by solving the fixed-point iteration

$$\begin{aligned} & 0 \in \gamma \nabla f(x_k) + \gamma \partial g(x_{k+1}) + x_{k+1} - x_k \\ \Leftrightarrow \quad & 0 \in \partial g(x_{k+1}) + \frac{1}{\gamma}\left(x_{k+1} - (x_k - \gamma \nabla f(x_k))\right) \\ \Leftrightarrow \quad & x_{k+1} = \mathbf{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right) \end{aligned}$$

and thus consists in alternating a proximal step and a gradient step.

3.3.2 Algorithm

The *proximal gradient* algorithm consists in iterating

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

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:

$$\begin{aligned} x_{k+1} &= \mathbf{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right) \\ &= \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\gamma}\|u - x_k + \gamma \nabla f(x_k)\|^2 \right\} \\ &= \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ g(u) + \langle u - x_k, \nabla f(x_k)\rangle + \frac{1}{2\gamma}\|u - x_k\|^2 + \frac{\gamma}{2}\|\nabla f(x_k)\|^2 \right\} \\ &= \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ f(x_k) + \langle u - x_k, \nabla f(x_k)\rangle + g(u) + \frac{1}{2\gamma}\|u - x_k\|^2 \right\} \end{aligned} \tag{3.4}$$

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 3.11. *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 (3.4), 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 1.14), 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 3.6). □

Bibliography

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

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.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

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

Claude Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 251(254):10, 2012.

Boris S Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*, volume 330. Springer Science & Business Media, 2006.

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 Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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

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

TUTORIAL 1 AROUND THE GRADIENT

Exercise 1.1 (Quadratic functions)**.**

a. In $\mathbb{R}^n$, compute the gradient of the squared Euclidean norm $\|\cdot\|_2^2$ at a generic point $x \in \mathbb{R}^n$.

b. Let A be an $m \times n$ real matrix and b a size-m real vector. We define $f(x) = \|Ax - b\|_2^2$. For a generic vector $a \in \mathbb{R}^n$, compute the gradient $\nabla f(a)$ and Hessian $\nabla^2 f(a)$.

c. Let C be an $n \times n$ real matrix, d a size-n real vector, and $e \in \mathbb{R}$. We define $g(x) = x^\top C x + d^\top x + e$. For a generic vector $a \in \mathbb{R}^n$, compute the gradient $\nabla g(a)$ and Hessian $H_g(a)$.

d. Can all functions of the form of f and be written in the form of g? And conversely?

Exercise 1.2 (Basic Differential calculus)**.** Use the composition lemma to compute the gradients of:

a. $f_1(x) = \|Ax - b\|_2^2$.

b. $f_2(x) = \|x\|_2$.

Exercise 1.3 (Optimality conditions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and $\bar{x} \in \mathbb{R}^n$. We suppose that f admits a local minimum at $\bar{x}$ that is $f(x) \geq f(\bar{x})$ for all x in a neighborhood[23] of $\bar{x}$.

a. For any direction $u \in \mathbb{R}^n$, we define the $\mathbb{R} \to \mathbb{R}$ function $q(t) = f(\bar{x} + tu)$. Compute $q'(t)$.

b. By using the first order Taylor expansion of q at 0, show that $\nabla f(\bar{x}) = 0$.

c. Compute $q''(t)$. By using the second order Taylor expansion of q at 0, show that $\nabla^2 f(\bar{x})$ is positive semi-definite.

[23] Formally, one would write $\forall x \in \mathbb{R}^n$ such that $\|x - \bar{x}\| \leq \varepsilon$ for $\varepsilon > 0$ and some norm $\|\cdot\|$.

Exercise 1.4 (Descent lemma)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-smooth if it is differentiable and its gradient ∇f is L-Lipchitz continuous, that is

$$\forall x, y \in \mathbb{R}^n, \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

The goal of the exercise is to prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then for all $x, y \in \mathbb{R}^n$,

$$f(x) \leq f(y) + (x - y)^\top \nabla f(y) + \frac{L}{2}\|x - y\|^2$$

a. Starting from fundamental theorem of calculus stating that for all $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) = \int_0^1 (x-y)^{\mathrm{T}} \nabla f(y + t(x-y)) \mathrm{d}t$$

prove the descent lemma.

b. Give a function for which the inequality is tight and one for which it is not.

Exercise 1.5 (Smooth functions)**.** Consider the constant stepsize gradient algorithm $x_{k+1} = x_k - \gamma \nabla f(x_k)$ on an L-smooth function f with some minimizer (i.e. some $x^\star$ such that $f(x) \geq f(x^\star)$ for all x).

a. Use the *descent lemma* to prove convergence of the sequence $(f(x_k))$ when $\gamma \leq 2/L$.

b. Does the sequence (x_k) converge? To what?

TUTORIAL 2 CONVEXITY

Exercise 2.1 (Fundamentals of convexity)**.**

a. Let f and g be two convex functions. Show that $m(x) = \max(f(x), g(x))$ is convex.

b. Show that $f_1(x) = \max(x^2 - 1, 0)$ is convex.

c. Let f be a convex function and g be a convex, non-decreasing function. Show that $c(x) = g(f(x))$ is convex.

d. Show that $f_2(x) = \exp(x^2)$ is convex. What about $f_3(x) = \exp(-x^2)$

e. Justify why the 1-norm, the 2 norm, and the squared 2-norm are convex.

Exercise 2.2 (Proof of Theorem 2.7)**.** Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. We want to show that if x is a local minimizer of F, then it is a (global) minimizer.

a. Since x is a local minimizer, 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 $\alpha = 1 - \rho/\|y - x\|$, $z = \alpha x + (1 - \alpha)y$. Show that $\alpha \in (0, 1)$ and $z \in \mathcal{B}(x, \rho)$.

b. Use convexity of F to conclude.

Exercise 2.3 (Strict and strong convexity)**.**

a. For a strictly convex function f, show that the problem

$$\begin{cases} \min f(x) \\ x \in C \end{cases}$$

where C is a convex set admits at most one solution.

b. Find a strictly convex function that admits no minimizer.

c. Show that a strongly convex function is also strictly convex.

Exercise 2.4 (Convexity and smoothness)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth convex function.

a. Show that for all $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) \leq \langle x - y; \nabla f(x) \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$$

and thus

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y; \nabla f(x) - \nabla f(y)\rangle \leq L\|x - y\|^2.$$

Hint: Define $z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x))$.
Use convexity to bound $f(x) - f(z)$ *and smoothness to bound* $f(z) - f(y)$ *and sum both inequalities.*

b. Let f be in addition μ-strongly convex with minimizer $x^\star$. Show that for all $x \in \mathbb{R}^n$,

$$(x - x^\star)^{\mathrm{T}}\nabla f(x) \geq \frac{\mu L}{\mu + L}\|x - x^\star\|^2 + \frac{1}{\mu + L}\|\nabla f(x)\|^2.$$

Hint: Use the fact that $f - \frac{\mu}{2}\|\cdot\|^2$ *is* $(L - \mu)$*-smooth and question a.*

TUTORIAL 3 THE PROXIMITY OPERATOR

Exercise 3.1 (Proof of Corollary 3.4)**.** Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function. Show that for any $y, z \in \mathbb{R}^n$

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

Exercise 3.2. Show that the proximity operator of the absolute value is given for $y \in \mathbb{R}$ and $\gamma > 0$ by

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

Exercise 3.3 (Application)**.** The *lasso* problem is a regularized linear regression problem that writes as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$$

where A is a full rank $m \times n$ matrix and b is a size m vector.

a. Write the iterations for a proximal gradient algorithm. Which stepsize can be used?

b. The regularization $\lambda\|x\|_1$ is said to be *sparsity enforcing*, guess why.

TUTORIAL 4 CONVERGENCE AND RATES

The goal of these exercises is to investigate the convergence and rate of the gradient algorithm on L-smooth functions.

Exercise 4.1 (Rate for smooth functions)**.** Let us consider a L-smooth convex function f and note $x^\star$ one of its minimizers. We consider the algorithm:

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k).$$

a. Prove that

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 - \frac{1}{L^2}\|\nabla f(x_k)\|^2 = \|x_k - x^\star\|^2 - \|x_{k+1} - x_k\|^2.$$

Hint: Replace x_{k+1} by $x_k - \frac{1}{L}\nabla f(x_k)$ and that for convex smooth functions $\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y; \nabla f(x) - \nabla f(y)\rangle$ as demonstrated in Exercise 2.4.

b. Show that

$$\delta_k := f(x_k) - f(x^\star) \leq \|x_k - x^\star\| \cdot \|\nabla f(x_k)\| \leq \|x_1 - x^\star\| \cdot \|\nabla f(x_k)\|.$$

Hint: Use convexity then a.

c. Use smoothness and b. to show that

$$0 \leq \delta_{k+1} \leq \delta_k - \underbrace{\frac{1}{2L\|x_1 - x^\star\|^2}}_{:=\omega}\delta_k^2.$$

d. Deduce that

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \omega.$$

Hint: Divide c. by $\delta_k\delta_{k+1}$.

e. Conclude that for the gradient algorithm with stepsize $1/L$ we have

$$f(x_k) - f(x^\star) \leq \frac{2L\|x_1 - x^\star\|^2}{k - 1}.$$

Exercise 4.2 (Iterates converence for smooth functions)**.** Under the same setup as the exercise above, we want to show that the iterates converge to some minimizer.

a. Show that the iterates are bounded and that for any $\overline{x} \in \operatorname{argmin} f$, $\|x_k - \overline{x}\|$ converges.

b. Show that every limit point of (x_k) belongs to $\operatorname{argmin} f$.

c. Suppose that $\overline{x}$ and $\overline{y}$ are two limit points of (x_k). Show that

$$2\langle x_k, \overline{x} - \overline{y}\rangle = \|x_k - \overline{y}\|^2 - \|x_k - \overline{x}\|^2 + \|\overline{x}\|^2 - \|\overline{y}\|^2$$

converges to some ℓ.

d. Using that $\overline{x}$ and $\overline{y}$ are two limit points of (x_k), deduce that $\ell = 2\langle \overline{x}, \overline{x} - \overline{y}\rangle = 2\langle \overline{y}, \overline{x} - \overline{y}\rangle$.

e. Deduce that (x_k) converges to a point in $\operatorname{argmin} f$.

Exercise 4.3 (Rate for smooth strongly convex functions)**.** Let us consider a L-smooth μ-strongly convex function f and note $x^\star$ its unique minimizer. We consider the algorithm:

$$x_{k+1} = x_k - \frac{2}{\mu + L}\nabla f(x_k).$$

a. Using Exercise 2.4, prove that

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &\leq \left(1 - \frac{4\mu L}{(\mu + L)^2}\right)\|x_k - x^\star\|^2 \\ &= \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \|x_k - x^\star\|^2\end{aligned}$$

where $\kappa = L/\mu$ is the *conditionning number* of the problem.

b. Show that

$$f(x_k) - f(x^\star) \leq \frac{L}{2}\|x_k - x^\star\|^2.$$

c. Conclude that for the gradient algorithm with stepsize $2/(\mu + L)$ we have

$$f(x_k) - f(x^\star) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \frac{L\|x_0 - x^\star\|^2}{2}.$$

d. Do the iterates converge?

TUTORIAL 5 LINEAR AND QUADRATIC PROGRAMS

In this tutorial, we are going to investigate Linear and Quadratic programs, often abbreviated LP and QP respectively. These problems appear when minimizing linear or quadratic cost functions under linear inequalities constraints. Typical formulations of these problems are:

Linear program (LP):

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^\top x \\ \text{subject to} \quad & Gx \leq h \end{aligned}$$

where $c, q \in \mathbb{R}^n$, $G \in \mathbb{R}^{m \times n}$, $h \in \mathbb{R}^m$, $P \in \mathbb{R}^{n \times n}$.

Quadratic program (QP):

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^\top P x + q^\top x \\ \text{subject to} \quad & Gx \leq h \end{aligned}$$

Although these problems are quite specific, a number of (sub-)problems in optimization can actually be formulated in one of the two forms above. The interest of these formulations is that there exists a large number of standard libraries implementing computationally efficient LP and QP solvers[24]. Depending on the solver, the formulation may vary but very marginally (for instance, some include linear equalities).

These solvers will take the vectors/matrices defining the problem and output a solution along with a solver status and additional information (precision, number of iterations, etc.). For instance, the library CVXOPT's functions are:
`lp(c, G, h)` and `qp(P, q, G, h)`
(see http://cvxopt.org/userguide/coneprog.html#linear-programming).

The main difficulty here is to properly reformulate the problems to make them fit the standard form.

[24] generally based on interior point, active sets, simplex, ... algorithms and variants.

Exercise 5.1 (First steps).

a. Do LP/QP always have solutions? Try to devise some conditions for solutions to exist if that is not the case.

b. My problem looks like a LP but the objective is $c^\top x + a$ where a is a constant scalar. Is that a problem?

c. My problem is $\min_{u,v} u + v$ under the constraints that $v \geq 0$ and $u = 7v + 3$. How can I formulate it as a LP?

Exercise 5.2 (Equivalent problems)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$; we consider the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & x \in C \end{aligned}$$

and we assume that a solution $\bar{x}$ exists. Show that this problem is *equivalent* to solving

$$\begin{aligned} \min_{(x,r) \in \mathbb{R}^{n+1}} \quad & r \\ \text{subject to} \quad & f(x) \leq r \\ & (x, r) \in C \times \mathbb{R} \subset \mathbb{R}^{n+1} \end{aligned}$$

in the sense that

(i) if $\bar{x}$ is a solution of the first problem, then $(\bar{x}, f(\bar{x}))$ is a solution of the second.

(ii) if $(\bar{x}, \bar{r})$ is a solution of the second problem, then $\bar{x}$ is a solution of the first one.

Exercise 5.3 (Linear reformulation)**.** Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Reformulate the problem

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|_\infty$$

as a linear problem. Notably, give the corresponding (c, G, h) from the LP formulation.

Exercise 5.4 (Linear reformulation II)**.** Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Reformulate the problem

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|_1$$

as a linear problem by extending the technique of Ex. 5.2 (without giving details). Notably, give the corresponding (c, G, h) from the LP formulation.

Do the same for the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_1 \\ \text{subject to} \quad & \|Ax - b\|_\infty \leq 1 \end{aligned}$$

[25] Drawings may be helpful!

Exercise 5.5 (Particular LPs)**.** Solve explicitly the following optimization problems:[25]

a. Minimize $c^\top x$ under the constraints that $\sum_i x_i = 1$ and $x \geq 0$ (linear function over the simplex).

b. Minimize $c^\top x$ under the constraint that $Ax = b$ (linear function over an affine subspace).

Exercise 5.6 (Quadratic reformulation)**.** We consider the regression model

$$y = X\theta + \xi, \quad \xi \sim \mathcal{N}(0, \sigma I_m),$$

where $X \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ are the observed values and $\theta \in \mathbb{R}^n$ is the unknown parameter we want to find.

a. Show that maximizing the likelihood $\ell(\theta) = f_\theta(y)$[26] of θ amount to minimizing $\|X\theta - y\|_2^2$.

b. Reformulate this problem as a QP.

[26] $f_\theta(y)$ is the density function of y for a fixed value of θ.