4 Convexity

Parts marked with ($\star$) are comments or more difficult questions. They are not discussed in the course and are not necessary for the evaluation.

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*.

4.1 Convex sets

4.1.1 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

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

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) is equivalent to

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

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. *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 (2), and thus of (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]

4.1.2 Convexity for sets

Let us now introduce the definition of a convex set.

Definition 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:

[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.

- 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"

4.1.3 Projection on convex sets

Getting back to the projection problem (1)

$$\min_{x \in C} F_y(x) \coloneqq \frac{1}{2}\|y - x\|^2 \tag{3}$$

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 (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 3. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, there is a unique projection* $\mathrm{proj}_C(y)$*, solution of (3).*

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

Theorem 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 [4, Th. 3.1.1]. □

4.2 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.

4.2.1 Definition

An extended real valued function[3] is convex if and only if its *epigraph*[4] is convex. However, the following definition is much more direct.

Definition 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 $\alpha \in (0, 1)$.

[3] A function that maps to $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

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

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 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. □

4.2.2 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 $\operatorname{dom} F \coloneqq \{x \in \mathbb{R}^n : F(x) < +\infty\}$, and its *infimum*

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

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

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

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

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

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.

4.2.3 (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 7 ([6, Prop. 8.12])**.** *Consider a convex proper lsc function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \operatorname{dom} F$. Then,*

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

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 8 ([2, Prop. 17.10]). *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 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.

4.2.4 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.

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

Theorem 9 ([6, Th. 6.12, 8.15]). *Consider a proper lower-semicontinuous convex function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a closed 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, equivalently,*

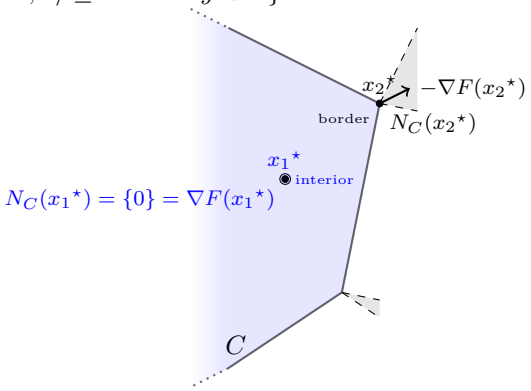
$$\langle y - x, v \rangle \geq 0$$

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 $y \in C$.

We recall that 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 $N_C(x) = \{0\}$.



4.2.5 Strict & strong convexity ($\star$)

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.

[5] *infeasible* in the optimization language.

Lemma 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 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 $v \in \partial F(x)$

$$F(u) \geq F(x) + \langle v, u - x\rangle + \frac{\mu}{2}\|u - x\|^2 \text{ for all } u \in \mathbb{R}^n \tag{4}$$

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 [2, Chap. 11.4] for a more general take).

Lemma 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. ($\star$) Let us consider the case where $C = \mathbb{R}^n$, the other cases can be deduced easily. From (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 [2, 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 to this compact set gives us the existence of a minimal value, which is unique from the quadratic lower bound expressed in (4). □

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

Theorem 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.

4.3 Smoothness and gradient descent

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.

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

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

for some $\gamma > 0$.

4.3.1 Smoothness

There is slight discrepancy in the literature concerning the notion of smoothness for functions. In [6], it is used for continuously differentiable functions, in Riemannian analysis it often refers to $\mathcal{C}^\infty$ function, while in numerical optimization and machine learning (see eg. [3]), 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 14. 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

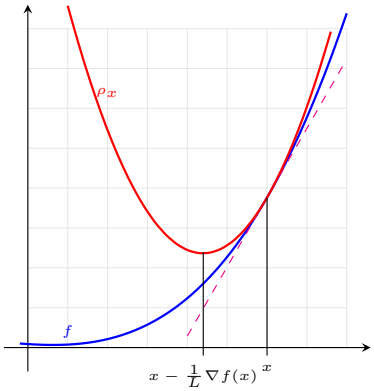
$$\|\nabla f(x) - \nabla f(u)\| \leq L\|x - u\| \text{ for all } x, u \in \mathbb{R}^n.$$

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

Lemma 15. *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 15 gives us

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

4.3.2 Gradient algorithm for convex functions

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

Theorem 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 (i.e., a point such that $\nabla f(x^\star) = 0$);*
- *(rate) $f(x_k) - f(x^\star) \leq \dfrac{2L\|x_0 - x^\star\|^2}{k}$ for any minimizer $x^\star$ of f.*

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 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 $\mathcal{O}(1/k^2)$; see [3, Th. 3.14]. This complexity is attained by Nesterov's fast gradient method [5]. 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 $\alpha_{k+1} = (k+2)/(k+3)$. (Actually, the choice for α_{k+1} is a bit more complicated but this variant grants the same rate.) ◀

4.3.3 Gradient algorithm for strongly convex functions (⋆)

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

Theorem 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 (unique by strong convexity);*
- *(rate) $f(x_k) - f(x^\star) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|x_0 - x^\star\|^2$ where $\kappa = \frac{L}{\mu} \geq 1$.*

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 $\mathcal{O}\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\right)$, again attained by a modified version of (Fast Gradient descent).

4.3.4 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:

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

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

This algorithm has similar guarantees as gradient descent.

Theorem 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 (that is 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 $y \in C$);*
- *(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.*

4.4 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.

4.4.1 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 20. *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 7 with $u = x^\star$ tells us that $g(x^\star) \geq g(x_k) + \langle v_k, x^\star - x_k\rangle$ 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,

$$\min_{\ell \leq k} g(x_\ell) - g(x^\star) \leq g\left(\frac{\sum_{\ell=0}^k \gamma_\ell x_\ell}{\sum_{\ell=0}^k \gamma_\ell}\right) - g(x^\star) \leq \frac{\sum_{\ell=0}^k \gamma_\ell (g(x_\ell) - g(x^\star))}{\sum_{\ell=0}^k \gamma_\ell}. \tag{5}$$

As for the right hand size:

a) if $\gamma_k = \gamma$, then

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

b) 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 (5). However, since the stepsize is decreasing, this limits the rate and the iterates convergence is out of reach. Nevertheless, its rate in $\mathcal{O}(1/\sqrt{k})$ is optimal on this class of functions.

4.4.2 Projected Subgradient algorithm ($\star$)

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

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

verifies the following properties.

Theorem 21. *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 $\sum_{k=0}^\infty \gamma_k^2 < +\infty$,

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

4.4.3 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

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

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

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

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

which is equivalent to saying that $x \mapsto L\frac{\|x\|^2}{2} - g(x)$ 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.

4.4.4 Non-Euclidean gradient descent ($\star$)

The simple yet powerful idea of [1] is then to compare g to a strictly convex function h in order to extend smoothness beyond the Euclidean case. Such h is usually called a *Bregman regularizer* or *distance-generating function* (DGF).

If there is an $L > 0$ such that $Lh - g$ is convex, or equivalently

$$(Lh(u) - g(u)) - (Lh(x) - g(x)) \leq \langle L\nabla h(x) - \nabla g(x), u - x\rangle,$$

then g is said *relatively smooth* wirh respect to h.

Then, we can define the associated *Bregman divergence* as

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

and use it to measure to distance between points.

The Euclidean gradient descent step can thus be transformed to

$$\begin{aligned} y &= \operatorname{argmin}_u \left\{ g(x) + \langle \nabla g(x); u - x\rangle + \frac{1}{2\gamma} D(u, x) \right\} \\ \Leftrightarrow \nabla h(y) &= \nabla h(x) - \gamma \nabla g(x) \end{aligned}$$

and some guarantees can be obtained in theory.

Exercises

Exercise 1. Let us consider the $\mathbb{R} \to \mathbb{R}$ function

$$F(x) = \frac{a}{2}x^2 + |x - 1|$$

for some $a > 0$.

1. Show that the function F is convex.
2. Compute the subgradient ∂F of F. Deduce the value of the minimizer of the function as a function of a.

Exercise 2. Let us consider the $\mathbb{R}^n \to \mathbb{R}$ function

$$G(x) = \frac{1}{2}\|x\|^2 + \iota_{\{x:\|x-e_1\|\leq\varepsilon\}}(x)$$

for some $\varepsilon > 0$, and where $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ elsewhere. The notation $\|\cdot\|$ stands for the Euclidean norm and $e_1 = [1, 0, \ldots, 0] \in \mathbb{R}^n$ is the first canonical vector.

1. Show that $\{x : \|x - e_1\| \leq \varepsilon\}$ is a closed convex set. Deduce that G is a convex proper lower semi-continuous function.
2. Show that $\operatorname{argmin}_{x\in\mathbb{R}^n} G(x) = \operatorname{argmin}_{x:\|x-e_1\|\leq\varepsilon} \|x\|^2$. What is the minimum of G in function of ε?

Exercise 3. Let us consider the $\mathbb{R} \to \mathbb{R}$ function

$$F(x) = \sum_{i=1}^{n} |x - i|$$

for some positive integer n.

1. Show that the function F is convex.
2. Compute the subgradient ∂F of F.
3. What are the minimizers of F depending on the value of n?
4. In the case when n is odd, is the function F strongly convex?

Exercise 4. Take a convex proper lower-semi-continuous $\mathbb{R} \to \mathbb{R}$ function F with a unique minimizer at point 1. Suppose that $x \mapsto F(x) + \varepsilon\, x^2$ also has a unique minimum at $x = 1$ for some $\varepsilon > 0$. Is it possible for F to be smooth?

References

[1] 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.

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

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

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

[5] 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.

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