INSTITUT NATIONAL
DES SCIENCES
APPLIQUÉES
TOULOUSE

Optimisation II

Optimisation lisse ou convexe: Théorie et algorithmie

Poly de cours

4A semestre 8

Dr. Frédéric de Gournay

Publié par INSA de Toulouse

frederic@degournay.fr

Première édition, Janvier 2024.

Table des matières

I Theory of smooth optimization

Theory of smooth optimization

Chapter 1

Refresher

We begin with some reminders which are prerequisites of this course of optimization.

1.1 First and second order differentiation

The difference bewteen **differentiation** and **derivative** must be bore in the mind of the student. One can talk about the latter only for functions from $\mathbb{R}$ to $\mathbb{R}$. If the function has several variables or yields several values, one cannot talk about **derivatives** but will rather talk about **differentiation**. The formal definition is as follows

Definition 1.1.1 Let $f : \mathbb{R}^n \to \mathbb{R}^p$. Then f is said to be **differentiable** at a point $x \in \mathbb{R}^n$ if there exists a linear mapping $d_x f : \mathbb{R}^n \to \mathbb{R}^p$, such that for every small step $h \in \mathbb{R}^n$, we have :

$$f(x+h) = f(x) + d_x f(h) + \mathcal{o}(\|h\|)\,.$$

This mapping $d_x f$ depends on x, it is called the **differential of f at the point x**.

The main fact that confuses the student is the **partial derivative** which is introduced for real-valued functions with several variables. The **partial derivative** is linked to the notion of **differential** (and to the idea of differentiation), but is not equivalent to it. It is important to understand the partial derivative but the goal of the analysis of functions of several variables is **differentiation**. The **partial derivative** is just a mean to that goal.

Definition 1.1.2 — Partial derivative. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ an let $1 \leq i \leq n$, we denote by $\frac{\partial f}{\partial x_i}(x)$ or $\partial_{x_i} f(x)$ or $\partial_i f(x)$ or even $f_{,i}(x)$, the **partial derivative** of f (if it exists) with respect to its i^{th} variable.

- This derivative is defined as the derivative at 0 of the function $t \mapsto f(x+te_i)$, where e_i is the i^{th} vector of the canonical basis.
- Equivalently, we have $\partial_i f(x) = \lim_{t\to 0} \frac{f(x+te_i)-f(x)}{t}$.
- This partial derivative is also sometimes called a **directional derivative** or **Gâteaux derivative**.

Exercice 1.1

Define $f : (x_1, x_2, x_3) \mapsto 3x_1^2 + 6x_1x_2 + x_3\cos(x_1x_2)$ and compute $\partial_2 f$.

Solution to Exercice 1.1

$\partial_2 f(x_1, x_2, x_3) = 6x_1 - x_3x_1\sin(x_1x_2)$

For any function from $\mathbb{R}^n \mapsto \mathbb{R}^p$, we have $n \times p$ different partial derivatives. They can be put in an array of size (p, n) and form a matrix. This matrix is coined as the **Jacobian**. The difficulty is to remember in which order the derivatives must be put (in column or in row ?). In other words, is the **Jacobian** a $n \times p$ or a $p \times n$ matrix ? For now, we must accept the following convention.

Definition 1.1.3 — Jacobian. Let $f : \mathbb{R}^n \mapsto \mathbb{R}^p$ and denote $f(x) = (f^1(x), \dots, f^p(x))$. Then the **Jacobian** of f at point $x \in \mathbb{R}^n$ is the $n \times p$ matrix given by:

$$Jac_x[f] = \begin{pmatrix} \partial_1 f^1(x) & \partial_2 f^1(x) & \dots & \partial_n f^1(x) \\ \partial_1 f^2(x) & \partial_2 f^2(x) & \dots & \partial_n f^2(x) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f^p(x) & \partial_2 f^p(x) & \dots & \partial_n f^p(x) \end{pmatrix}$$

Exercice 1.2

Compute the Jacobian of $f : \mathbb{R}^3 \mapsto \mathbb{R}$ if $f(x) = \begin{pmatrix} 3x_1^2 + x_3\cos(x_1x_2) \\ 5x_2x_1 + \ln(x_3^2 + 2) \end{pmatrix}$

Solution to Exercice 1.2

$$Jac_x[f] = \begin{pmatrix} 6x_1 - x_2x_3\sin(x_1x_2) & -x_1x_3\sin(x_1x_2) & \cos(x_1x_2) \\ 5x_2 & 5x_1 & \frac{2x_3}{x_3^2+2} \end{pmatrix}$$

Theorem 1.1.1 Let $f : X \to \mathbb{R}^{|}$, where $X \subset \mathbb{R}^n$ and suppose that f is differentiable at point $x \in X$, then the Jacobian is the matrix of the differential in the canonical basis. Hence, for any $h \in X$, we have

$$d_x f(h) = (Jac_x[f]).h$$

Finally, in the case of a real-valued function (and only in this case), we can define the **gradient**.

Definition 1.1.4 — Gradient. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$, suppose that the Jacobian of f at point x exists. Then its transpose is denoted $\nabla f(x)$ and called the **gradient**. We then have:

$$\nabla f(x) = (Jac_x[f])^T = \begin{pmatrix} \partial_1 f(x) \\ \partial_2 f(x) \\ \vdots \\ \partial_n f(x) \end{pmatrix}$$

Exercice 1.3

If $x = (x_1, x_2)$ and $f(x) = 3x_1^2 + x_2\cos(x_1)$, compute the gradient of f.

Solution to Exercice 1.3

$$\nabla f(x_1, x_2) = \begin{pmatrix} 6x_1 - x_2\sin(x_1) \\ \cos(x_1) \end{pmatrix}$$

The relationship between the Jacobian, the gradient and the differential is as follows.

Theorem 1.1.2 Let $f : X \to \mathbb{R}$, where $X \subset \mathbb{R}^d$ and suppose that f is differentiable at point $x \in X$. The gradient of f at point x is the only vector such that

$$\langle \nabla f(x), h \rangle = d_x f(h) \quad \forall h$$

In optimization, second derivatives matter. Writing down second order differentiation can be quite messy, but is simpler in the case where the function is real-valued.

Definition 1.1.5 Let $f : \mathbb{R}^n \mapsto \mathbb{R}$, then we call **Hessian** of f at point x the $n \times n$ matrix of second-order derivative

$$H[f](x) = Jac_x[\nabla f] = \begin{pmatrix} \partial_{11} f(x) & \partial_{12} f(x) & \dots & \partial_{1n} f(x) \\ \partial_{21} f(x) & \partial_{22} f(x) & \dots & \partial_{2n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{n1} f(x) & \partial_{n2} f(x) & \dots & \partial_{nn} f(x) \end{pmatrix}$$

Exercice 1.4

Compute the gradient and the Hessian of

$$f : x = (x_1, x_2) \mapsto 3x_1^2 + x_2\cos(x_1),$$

Solution to Exercice 1.4

$$\nabla f(x) = \begin{pmatrix} 6x_1 - x_2\sin(x_1) \\ \cos(x_1) \end{pmatrix} \text{ et } H[f](x) = \begin{pmatrix} 6 - x_2\cos(x_1) & -\sin(x_1) \\ -\sin(x_1) & 0 \end{pmatrix}$$

For a regular function, immediate computations shows that the partial derivative of a function with respect to x and then to y is the same than the one with respect to y and then to x. In other words, the order in which the partial derivative are taken does not matter. This is actually equivalent to stating that the Hessian matrix is symetric, this theorem is known as **Schwart'z theorem**.

Proposition 1.1.3 — Schwartz. If f is a C^2 function, then $H[f](x) = (H[f](x))^T$

Shcwartz theorem is not always true, first there exists counterexamples of weird functions which admits second order partial derivative but are not C^2 and for which the order of directions of derivations matters. A more important counterexample is at the core idea of Riemanian geometry. What Schwartz theorem states is that taking a step in the x direction and then in the y direction is the same thing than taking a step in the y and then in the x direction. But this is true only on a plane, in everydays's life, this theorem is actually false. Indeed, on a sphere, if one takes a step in the east direction and then on the north direction, he does not end up in the same place than if he takes a step first towards the north and then the east. Schwart'z theorem is only true in the so-called euclidean or plane geometry. It fails to be true on curved or Riemanian geometry.

1.2 Taylor expansion

Taylor expansions are a major tool in mathematics and they are of utmost importance in optimization. We refresh the notion of Taylor expansion for functions from $\mathbb{R}$ to $\mathbb{R}$.

Proposition 1.2.1 — Classic Taylor expansions. If $f : \mathbb{R} \mapsto \mathbb{R}$ is sufficiently regular, we have:

$$\begin{array}{rcl} f(x+h) & = & f(x) + f'(x)h + o(h) \\ f(x+h) & = & f(x) + f'(x)h + \mathcal{O}(h^2) \\ f(x+h) & = & f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^2) \\ f(x+h) & = & f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \mathcal{O}(h^3) \end{array}$$

The general form is given by

$$f(x+h) = f(x) + f'(x)h + \cdots + \frac{f^{(n)}(x)}{n!}h^n + R,$$

where R is the remainder of the Taylor expansion, it can be given as

$$\begin{cases} R = o(h^n) & \text{if } f \text{ is } D^n \text{ (Peano remainder)} \\ R = \mathcal{O}(h^{n+1}) & \text{if } f \text{ is } C^{n+1} \\ R = \frac{f^{(n+1)}(x+\xi)}{(n+1)!}h^{n+1} \text{ for } 0 \le \xi \le h & \text{if } f \text{ is } D^{n+1} \text{ (Mean-value remainder)} \\ R = \int_0^h \frac{f^{(n+1)}(t)}{n!}(h-t)^n dt & \text{if } f^{(n+1)} \text{ is integrable (Integral remainder)} \end{cases}$$

The general form of the Taylor expansion at first order can be inferred from the unidimensional case.

Proposition 1.2.2 — General case. If $f : \mathbb{R}^n \mapsto \mathbb{R}^p$ is regular enough, we have :

$$f(x+h) = f(x) + Jac_x[f].h + o(\|h\|)$$

In the case of a real-valued function, the Taylor expansion must be known

Proposition 1.2.3 — Real valued function. If $f : \mathbb{R}^n \mapsto \mathbb{R}$, is regular enough, we have :

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

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle H[f](x)h, h \rangle + o(\|h\|^2)$$

1.3 Levelsets and gradients

If $f : \mathbb{R}^n \mapsto \mathbb{R}$, then for each point $M \in \mathbb{R}^n$, $\nabla f(M)$ is a vector of $\mathbb{R}^n$. In the case $n = 2$, the graph of a function is a $2d$ surface in a $3d$ plot but the gradient is a $2d$ vector. Hence it has to be plotted in the (x, y) plane. In Figure 1.2, we plot in 3 dimensions the graph of the function $f(x, y) = x^2 + 0.8 * y^2 + 0.7$, in yellow/ocre coulours, we plot the surface $z = x^2 + 0.8 * y^2 + 0.7$. We also plot in black three levelsets and their projection in the $z = 0$ plane. At the point $(x_0, y_0) = (0.4, 0.3)$, we plot the tangent space in black. It is given by the equation:

$$z = \left\langle \nabla f(x_0, y_0), \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right\rangle + f(x_0, y_0)$$

This equation can be rewritten

$$\left\langle \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} \nabla f(x_0, y_0) \\ -1 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} x_0 \\ y_0 \\ f(x_0, y_0) \end{pmatrix}, \begin{pmatrix} \nabla f(x_0, y_0) \\ -1 \end{pmatrix} \right\rangle.$$

Hence the vector $\begin{pmatrix} \nabla f(x_0, y_0) \\ -1 \end{pmatrix}$ is orthogonal to the tangent plane. It is displayed, attached to the point $(x_0, y_0, f(x_0, y_0))$ in blue in Figure 1.2. The vector $\nabla f(x_0, y_0)$ attached at the point (x_0, y_0) and is displayed in red in Figure 1.2.

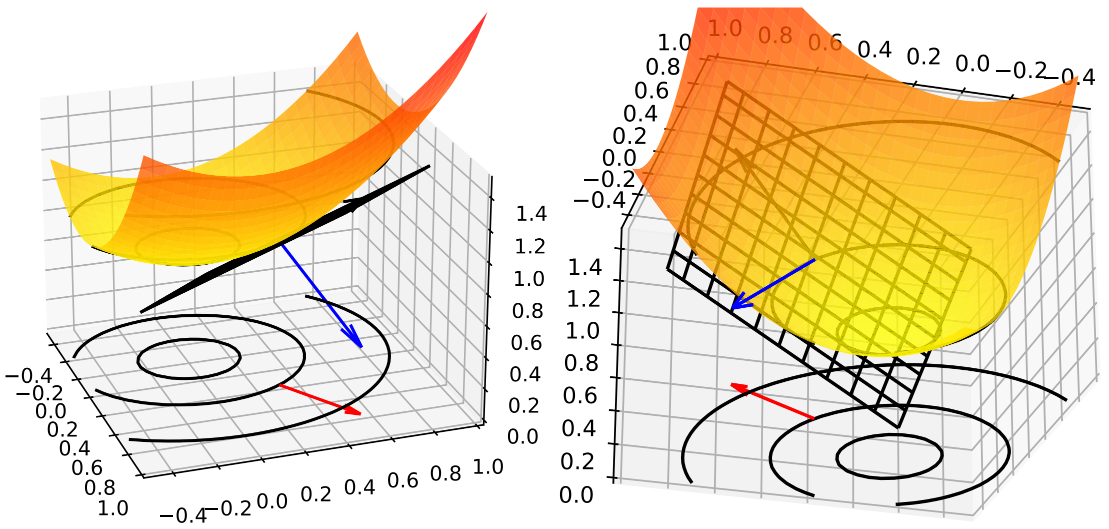


Figure 1.1: A quadratic function and its Taylor expansion in $(x_0, y_0) = (0.4, 0.3)$ its levelset are in black and its tangent plane at (x_0, y_0) is in black. In red, the gradient at (x_0, y_0) , in blue the vector $\begin{pmatrix} \nabla f(x_0, y_0) \\ -1 \end{pmatrix}$, in black the vector $\begin{pmatrix} \nabla f(x_0, y_0) \\ \|\nabla f(x_0, y_0)\|^2 \end{pmatrix}$.

1.4 Symetric matrices

Before being able to state the main theorem, we need some preliminary definitions and technical lemmatas.

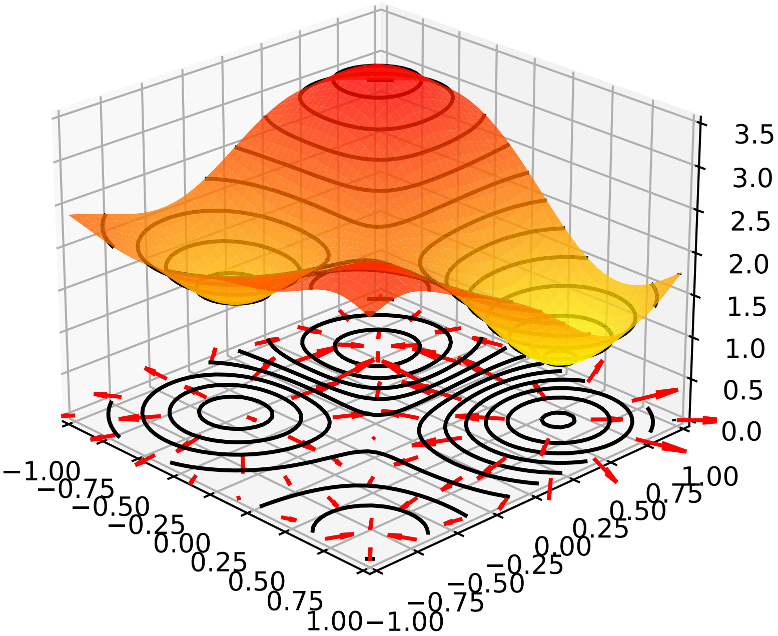


Figure 1.2: A function, its levelsets and the gradient field. Notice that the gradient is orthogonal to the level-sets.

Definition 1.4.1 Let A be a symetric matrix in $\mathcal{M}_{n\times n}(\mathbb{R})$. Then all the eigenvalues of A are real and A is diagonalizable in an orthonormalized basis.

- If every eigenvalue of A is non-negative ($i.e. \geq 0$), we say that A is semi-definite positive and we denote $A \succeq 0$.
- If every eigenvalue of A is positive ($i.e. > 0$), we say that A is definite positive and we denote $A \succ 0$.

The fact that all the eigenvalues of A are real and that there exists an orthonormal basis that diagonalizes A is supposed to be known to the reader. We have a caracterization of semi-definite positive and definite positive matrices. It reads :

Proposition 1.4.1 Let A be a symetric matrix in $\mathcal{M}_{n\times n}(\mathbb{R})$. Then,

$$A \succeq 0 \iff \langle Ah, h\rangle \geq 0 \quad \forall h \in \mathbb{R}^n \iff \exists c \geq 0 \text{ s.t. } \langle Ah, h\rangle \geq c\|h\|^2 \quad \forall h \in \mathbb{R}^n.$$

For the definite positive case, the proposition is almost the same. We assume that $h \in \mathbb{R}^n$.

$$A \succ 0 \iff \langle Ah, h\rangle > 0 \quad \forall h \neq 0 \iff \exists c > 0 \text{ s.t. } \langle Ah, h\rangle \geq c\|h\|^2 \quad \forall h$$

Proof

Let (e_i) be an orthonormalized basis that diagonalizes A and $(\lambda_i)_i$ be the corresponding eigenvalues. The existence of such a basis is ensured by the fact that A is a symmetric matrix with real coefficients. By definition we have

$$(Ae_i = \lambda_i e_i \quad \forall i) \quad \text{and} \quad (\langle e_i, e_j\rangle = 0 \quad \forall i \neq j) \quad \text{and} \quad (\langle e_i, e_i\rangle = 1 \quad \forall i)$$

Let $h \in \mathbb{R}^n$, decompose h on the basis $(e_i)_i$ and denote $(h_i)_i$ the coordinates, we have

$$h = \sum_i h_i e_i.$$

We have the standard Pythagorean equality

$$\langle h, h \rangle \;=\; \left\langle \left(\sum_{i=1}^n h_i e_i \right), \left(\sum_{j=1}^n h_j e_j \right) \right\rangle = \sum_{i,j} h_i h_j \langle e_i, e_j \rangle = \sum_{i=1}^n h_i^2.$$

Denote λ_m the smallest eigenvalue of A, it comes

$$\begin{aligned} \langle Ah, h \rangle &= \left\langle A \left(\sum_{i=1}^n h_i e_i \right), \left(\sum_{j=1}^n h_j e_j \right) \right\rangle = \sum_{i,j} h_i h_j \langle A e_i, e_j \rangle = \sum_{i,j} h_i h_j \langle \lambda_i e_i, e_j \rangle \\ &= \sum_{i,j} \lambda_i h_i h_j \langle e_i, e_j \rangle = \sum_{i=1}^n \lambda_i h_i^2 \geq \sum_{i=1}^n \lambda_m h_i^2 \geq \lambda_m \|h\|^2, \end{aligned}$$

We now prove the equivalences

- Suppose that $A > 0$ (resp. ≥ 0), then $\lambda_m > 0$ (resp. ≥ 0) and for any h, we have
$$\langle Ah, h \rangle \geq \lambda_m \|h\|^2$$
- Suppose that there exists $c > 0$ (resp. ≥ 0) such that
$$\langle Ah, h \rangle \geq c\|h\|^2,$$
then $\langle Ah, h \rangle > 0$ (resp. ≥ 0) for any $h \neq 0$.
- Suppose that $\langle Ah, h \rangle > 0$ (resp. ≥ 0) for any $h \neq 0$. By choosing $h = e_i$, we have $\langle Ah, h \rangle = \lambda_i$. So that every eigenvalue of A is > 0 (resp. ≥ 0) and hence $A \geq 0$ (resp. > 0).

1.5 Convexity of sets

1.5.1 Definition

Definition 1.5.1 — Set convexity. Let E be any vector space and $\mathcal{C} \subset E$, we say that $\mathcal{C}$ is **convex** if and only if we have

$$\forall (x, y) \in \mathcal{C}^2, \quad \forall \theta \in]0, 1[, \text{ then } \quad \theta x + (1 - \theta) y \in \mathcal{C}$$

In words, a set $\mathcal{C}$ is convex if and only if it contains every segment that ends in $\mathcal{C}$. In Figure 1.3, we display several examples of convex sets. In Figure 1.4, we emphasize the fact that the boundary is of importance when considering convexity. Indeed, in Figure 1.4, we represent four sets which have same interior and same closure, only the first three of those sets are convex.

The following proposition states that convex sets are stable by intersection. This property is of importance because it allows to talk about the smallest convex set in the sense of the inclusion.

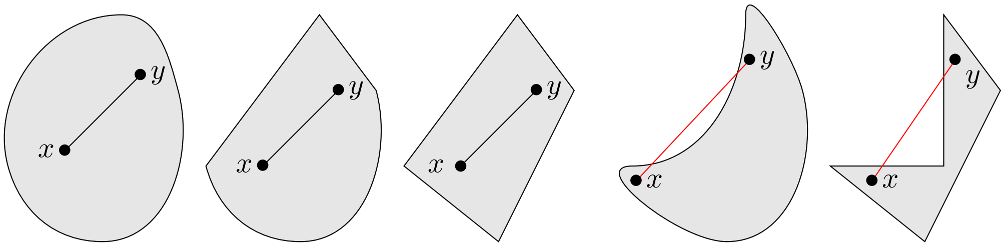


Figure 1.3: Convexity of sets, the first three sets are convex while the last two are not.

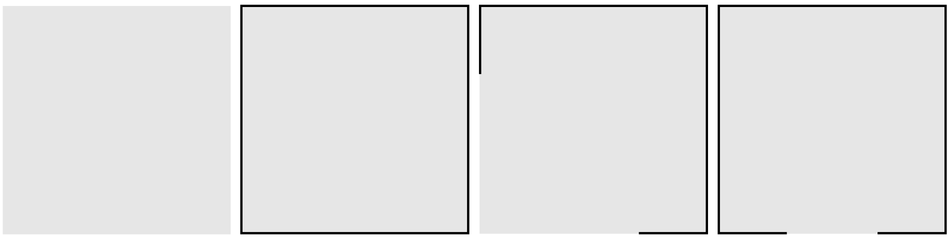


Figure 1.4: Convexity of sets, the first three sets are convex while the last one is not. In black is the part of the boundary that belongs to the set.

Proposition 1.5.1 Let $(\mathcal{C}_i)_{i \in I}$ be a family of convex sets, then $\mathcal{C}^\star = \bigcap_{i \in I} \mathcal{C}_i$ is a convex set. For any property (P) which is stable by intersection, we can talk about the the smallest convex set that verifies (P).

Proof

- Let x and y belong to $\mathcal{C}^\star$ and take $\theta \in]0,1[$. For any $i \in I$, both x and y belong to $\mathcal{C}_i$, so that $\theta x + (1-\theta)y$ belongs to $\mathcal{C}_i$. Hence $\theta x + (1-\theta)y$ belongs to $\mathcal{C}^\star$ and therefore $\mathcal{C}^\star$ is convex.
- Take any property (P), take I the family of convex sets that verifies (P) and $\mathcal{C}^\star$ the intersection of all the sets in I. If (P) is stable by intersection, then $C^\star$ verifies (P) and $\mathcal{C}^\star$ is included in every sets that contains (P). Hence $\mathcal{C}^\star$ the smallest set that verifies (P).

For instance, for any given set A, we can talk about the smallest set that contains A, and we can give it a name (spoiler, it is called the convex hull).

A very important class of convex sets is the class of polytopes.

Proposition 1.5.2 — Polytopes. Let $A \in \mathcal{M}_{p,n}(\mathbb{R})$, $C \in \mathcal{M}_{q,n}(\mathbb{R})$, $b \in \mathbb{R}^p$ and $d \in \mathbb{R}^q$. Define $\mathcal{C} \subset \mathbb{R}^n$ as

$$\mathcal{C} = \{x \in \mathbb{R}^n \mid Ax \preceq b, Cx = d\}.$$

Here $\mathcal{C}$ is defined with a finite number of equalities and inequalities. The set $\mathcal{C}$ is a convex set and is called a polytope. In dimension 2, the polytopes are called the polygons and in dimension 3, they are called the polyhedra (the singular is polyhedron).

1.5.2 Convex Hull

Definition 1.5.2 — Convex combination. Let $x_1, .., x_m \in \mathbb{R}^n$. We say that x is a convex combination of $(x_i)_{1 \leq i \leq m}$ if there exists real **non-negative** coefficients $(\alpha_1, ..., \alpha_m)$ such that :

$$x = \sum_{i=1}^{m} \alpha_i x_i \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i = 1$$

In other words, a convex combination is an **ponderated average** with non-negative coefficients. In Figure 1.5, we display some examples of family $(x_i)_{i \in I}$ and the corresponding convex combinations.

Definition 1.5.3 — Convex hull. Let $X \subseteq \mathbb{R}^n$, the convex hull of X is denoted conv (X) and is defined as the smallest convex set that contains X. In finite dimension, the convex hull of X is the set of convex combinations of elements of X :

$$\operatorname{conv}(X) = \{x \in \mathbb{R}^n \mid x = \sum_{i=1}^{p} \alpha_i x_i \text{ where } x_i \in X,\ p \in \mathbb{N} \text{ and } \sum_{i=1}^{p} \alpha_i = 1,\ \alpha_i \geq 0\}.$$

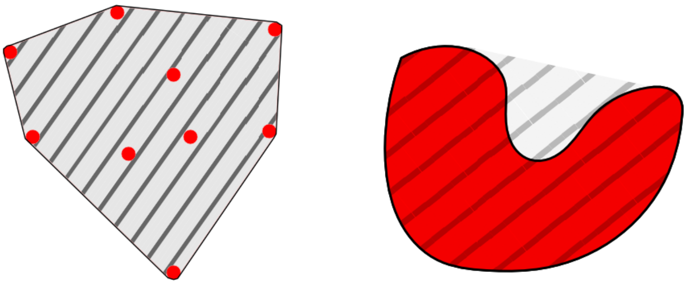


Figure 1.5: Exemples of convex hulls. On the left, a typical convex hull of a finite number of points. On the right, a convex hull of a domain with an infinite number of points.

1.6 Convexity of functions

We turn our attention to the notion of convexity for a function. In a nutshell, a function f is convex if and only if **f of the average is smaller than the average of the f's**. In this section, we consider functions from X to $\mathbb{R}$, where X is a convex set. These functions are called **real-valued** functions.

Definition 1.6.1 — Real-valued convex function. Let $f : X \mapsto \mathbb{R}$ with X a convex set. We say that f is **convex** over X if and only if

$$\forall (x, y) \in X^2, \quad \forall \theta \in]0, 1[, \quad f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

We say that f is **strictly convex** over X if and only if

$$\forall (x, y) \in X^2, \quad \forall \theta \in]0, 1[, \quad f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$$

An illustration is given in 1.6. In this figure, we see that a convex function may fail to be differentiable at some points. We can prove however that a real-valued convex function is almost everywhere differentiable on X.

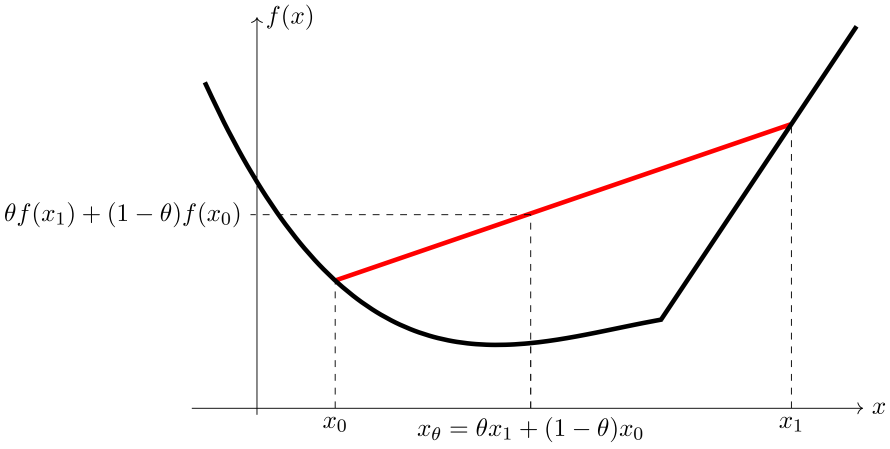


Figure 1.6: An example of convex function from $\mathbb{R}$ to $\mathbb{R}$. Note that this function is not differentiable everywhere

In Figure 1.6, the segment between $(x, f(x))$ and $(y, f(y))$ is called **a chord**. By Definition 1.6.1, convex function lies under their chords. It turns out that convex functions lies above their tangents (Taylor expansion of first order). This is made clear in the following proposition. An illustration of this fact is given on Figure 1.7 in $2d$ and $3d$.

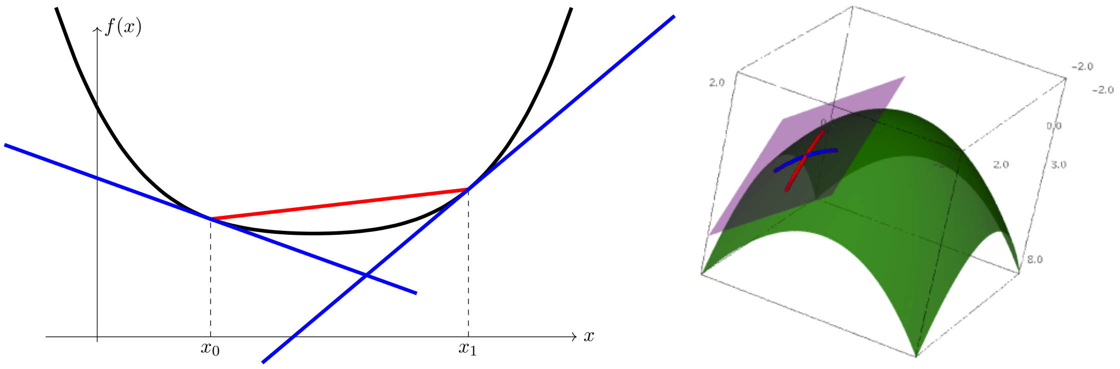


Figure 1.7: A convex function is always above its tangent hyperplanes (when they exists), on the left a $1d$ example and on the right a $2d$ example.

Theorem 1.6.1 Let $X \subset \mathbb{R}^n$ be a convex set and $f : X \to \mathbb{R}$ be differentiable on X. The function f is convex over X if and only if

$$\forall (x, y) \in X^2, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \tag{1.1}$$

Proof

We first suppose that f is convex. Let x and y be any element of X, by convexity of f, it holds for every $\theta \in]0,1[$ that :

$$f((1-\theta)x+\theta y) \leq (1-\theta)f(x)+\theta f(y) = f(x)+\theta(f(y)-f(x)).$$

Reorganizing the above inequality, we obtain:

$$\frac{f(x+\theta(y-x))-f(x)}{\theta} \leq f(y)-f(x).$$

Letting θ go to zero while being positive, we obtain (1.1). We now suppose that (1.1) is true. For any $(a,b) \in X^2$ and $\theta \in [0,1]$, apply (1.1) to $(x=\theta a+(1-\theta)b, y=a)$, and then to $(x=\theta a+(1-\theta)b, y=b)$, we obtain

$$\begin{array}{ccccccc} f(a) & \geq & f(\theta a+(1-\theta)b) & + & (1-\theta) & \langle \nabla f(\theta a+(1-\theta)b), b-a\rangle \\ f(b) & \geq & f(\theta a+(1-\theta)b) & - & \theta & \langle \nabla f(\theta a+(1-\theta)b), b-a\rangle \end{array}$$

Multiply the first inequality by θ and the second by $1-\theta$. Add up the two inequalities and obtain:

$$\theta f(a)+(1-\theta)f(b) \geq f(\theta a+(1-\theta)b),$$

Hence f is convex on X.

We stated that a function is convex if and only if **f of the average is smaller than the average of the f's**. But this statement is true in definition 1.6.1 only when taking the average of two points x and y. In fact, this statement is true when taking the average (we rather talk about a **ponderated barycenter** in this case) of any finite number of point. This theorem is known as Jensen's inequality.

Proposition 1.6.2 — Jensen's inequality. Let $f:\mathbb{R}^n \to \mathbb{R}$ be a convex function over X and $(x_i)_i$ a family of points of X and $(\alpha_i)_i \geq 0$ some real coefficients such that $\sum_{i=1}^m \alpha_i = 1$, then

$$f\left(\sum_{i=1}^m \alpha_i x_i\right) \leq \sum_{i=1}^m \alpha_i f(x_i)$$

Proof

The proof is by recurrence over m. For the case $m=2$, we must have $\alpha_2 = 1-\alpha_1$ and then

$$f(\alpha_1 x_1+(1-\alpha_1)x_2) \leq \alpha_1 f(x_1)+(1-\alpha_1)f(x_2).$$

Suppose now that the results holds true for $m \leq k$. Set $m=k+1$ and denote $\beta = \sum_{i=1}^k \alpha_i$ and $y = \sum_{i=1}^k \frac{\alpha_i}{\beta}x_i$. Apply the recurence hypothesis to the family $(x_i)_{1\leq i \leq k}$ with coefficients $(\frac{\alpha_i}{\beta})_{1\leq i\leq k}$. Note that the coefficients are positive and sum up to on. We have

$$f(y) = f(\sum_{i=1}^k \frac{\alpha_i}{\beta}x_i) \leq \frac{1}{\beta}\left(\sum_{i=1}^k \alpha_i f(x_i)\right)$$

Now use the above inequality with convexity (case $k=2$, with $\beta+\alpha_m=1$) to obtain

:

$$f\left(\sum_{i=1}^{m}\alpha_i x_i\right) = f(\beta y + \alpha_m x_m) \underbrace{\leq}_{\text{convexity}} \beta f(y) + \alpha_m f(x_m)$$
$$\leq \sum_{i=1}^{k}\alpha_i f(x_i) + \alpha_m f(x_m)$$

1.7 Proving convexity

It is rather difficult to prove convexity with the definitions of the previous section. Hopefully we have some usefull theorem that helps determining wether an object is convex or not

Theorem 1.7.1 — Convexity of sets.

1. If X is a vector space, then X is convex.
2. If X is defined via inequality and equality constraints, that is

$$X = \{x, g_i(x) \leq 0 \text{ and } h_j(x) = 0 \quad \forall i \in I \text{ and } j \in J\}.$$

If g_i is convex for every $i \in I$ and h_j is affine for every $j \in J$, then X is convexe

Proof

1. Let x and y be in a vector space, then for any $\theta \in [0,1]$, $\theta x + (1-\theta)y$ also belongs to the same vector space.
2. For each i and j, define

$$X_i = \{x, g_i(x) \leq 0\} \quad \text{ and } X_j = \{x, h_j(x) = 0\}.$$

If we show that each X_i and X_j are convex, because $X = (\cap_i X_i) \cap (\cap_j X_j)$, we are done.

- Let us begin by X_i. If x and y are in X_i and $\theta \in [0,1]$, then

$$g(\theta x + (1-\theta)y) \underbrace{\leq}_{\text{convexity of } g} \underbrace{\theta}_{\geq 0}\underbrace{g(x)}_{\leq 0} + \underbrace{(1-\theta)}_{\geq 0}\underbrace{g(y)}_{\leq 0} \leq 0$$

- Let x and y be in X_j and $\theta \in [0,1]$. Because h_j is affine, there exists a linear function a and $b \in \mathbb{R}$ such that $h_j(z) = a(z) + b$ for every z. We then have:

$$\begin{aligned} h(\theta x + (1-\theta)y) &= a(\theta x + (1-\theta)y) + b \\ &= \theta a(x) + (1-\theta)a(y) + (\theta + (1-\theta))b \\ &= \theta(a(x)+b) + (1-\theta)(a(y)+b) \\ &= \theta \underbrace{h(x)}_{=0} + (1-\theta)\underbrace{h(y)}_{=0} \\ &= 0 \end{aligned}$$

Theorem 1.7.2 — Necessary conditions for convex functions. Let f be a C^2 function over a convex set $X \subset \mathbb{R}^n$. If f is a convex function over X then $H[f](x) \succeq 0$ for every $x \in \overset{\circ}{X}$ (the interior of X).

Proof

Let f be a convex function and suppose there exists $x \in \mathbb{R}^n$ such that $H[f](x)$ admits a negative eigenvalue λ. Let d be an eigenvector associated to λ. Because $x \in \overset{\circ}{X}$, then $x + \varepsilon d \in X$ for every $\varepsilon > 0$ small enough. Moreover

$$\begin{aligned}
f(x+\varepsilon d) &= f(x) + \langle \nabla f(x), \varepsilon d \rangle + \frac{\varepsilon^2}{2}(H[f](x)d, d) + \mathcal{o}\big(\varepsilon^2\big) \\
&= f(x) + \langle \nabla f(x), \varepsilon d \rangle + \frac{\varepsilon^2}{2}\underbrace{\lambda \|d\|^2}_{<0} + \mathcal{o}\big(\varepsilon^2\big) \\
&< f(x) + \langle \nabla f(x), \varepsilon d \rangle \text{ for small enough } \varepsilon.
\end{aligned}$$

Hence if f is convex, there is a contradiction with Theorem 1.6.1.

Theorem 1.7.3 — Sufficient conditions for convex functions. If f is a C^2 function over a convex set $X \subset \mathbb{R}^n$, then

- If $H[f](x) \succeq 0$ for all $x \in X$, then f is convex on X .
- If $H[f](x) \succ 0$ for all $x \in X$, then f is strictly convex on X .

The proof of this result is divided into two steps. We first prove it in the $1d$ case $X = [0,1]$ and where we are only interested in the convexity between the points 0 and 1. That is, we prove the following lemma :

Lemma 1.7.4 Let $\phi \in \mathcal{C}^2([0,1])$ and $\theta \in]0,1[$.

- If $\phi''(x) > 0$ for all $x \in [0,1]$, then $\phi(\theta) < (1-\theta)\phi(0) + \theta\phi(1)$.
- If $\phi''(x) \geq 0$ for all $x \in [0,1]$, then $\phi(\theta) \leq (1-\theta)\phi(0) + \theta\phi(1)$.

Proof

We only deal with the case $\phi'' > 0$, for the case $\phi'' \geq 0$, replace strict inequalities by large one. Because ϕ' is increasing, we have

$$\begin{aligned}
\phi(\theta) - \phi(0) &= \int_0^\theta \phi'(t)dt < \theta\phi'(\theta) \\
\phi(1) - \phi(\theta) &= \int_\theta^1 \phi'(t)dt > (1-\theta)\phi'(\theta)
\end{aligned}$$

Grouping these two inequalities, we have

$$\frac{\phi(\theta)-\phi(0)}{\theta} < \phi'(\theta) < \frac{\phi(1)-\phi(\theta)}{1-\theta} \Rightarrow (1-\theta)(\phi(\theta)-\phi(0)) < \theta(\phi(1)-\phi(\theta))$$

which means: $\phi(\theta) < (1-\theta)\phi(0) + \theta\phi(1)$.

Proof of Theorem 1.7.3

Suppose $H[f](z) > 0$ for all $z \in X$. The case $H[f](z) \geq 0$ is done the same way and we do not prove this case. Let $x, y \in X^2$. We introduce the $\mathcal{C}^2$ function

$\phi : [0,1] \longrightarrow \mathbb{R}$ defined by

$$\phi(\theta) = f(\theta x + (1-\theta)y) \quad \forall \theta.$$

In order to compute the second derivative of ϕ, we use the second order Taylor expansion of f. For any θ and $\dot{\theta}$, we have

$$\phi(\theta + \dot{\theta}) = f((\theta + \dot{\theta})x + (1 - \theta - \dot{\theta})y) = f(\theta x + (1-\theta)y + \dot{\theta}(x-y)).$$

Introducing $m = \theta x + (1-\theta)y$, we have

$$\begin{aligned} \phi(\theta + \dot{\theta}) &= f(m + \dot{\theta}(x-y)) \\ &= f(m) + \dot{\theta}\underbrace{\langle \nabla f(m), (x-y)\rangle}_{(\mathbf{1})} + \frac{\dot{\theta}^2}{2}\underbrace{\langle H[f](m)(x-y), (x-y)\rangle}_{(\mathbf{2})} + o\left(\dot{\theta}^2\right) \end{aligned}$$

We can identify $(\mathbf{1})$ as the first order derivative of ϕ and $(\mathbf{2})$ as its second order derivative. We have

$$\begin{aligned} \phi'(\theta) &= \langle \nabla f(m), x-y\rangle, \\ \phi''(\theta) &= \langle H[f](m)(x-y), (x-y)\rangle. \end{aligned}$$

Because $H[f](m) > 0$ and $x \neq y$, then $\phi''(m) > 0$. We apply Lemma 1.7.4 to obtain:

$$f(\theta x + (1-\theta)y) < (1-\theta)f(y) + \theta f(x),$$

because: $\phi(0) = f(y)$, $\phi(1) = f(x)$ and $\phi(\theta) = f(\theta x + (1-\theta)y)$.
This proves that f is strictly convex.

1.8 Minimum, infimum, supremum and maximum

Definition 1.8.1 — Infimum of a set. Let X a subset of a Banach space and $f : X \mapsto \mathbb{R}$. The set of lower bounds of f over X is the set

$$\{z, \text{such that } z \leq f(x) \quad \forall x \in X\}$$

We define the **infimum** of f over X and denote it $\inf_X(f)$ as :

- If X is empty then $\inf_X(f) = +\infty$.
- If the set of lower bounds of f over X is empty, then $\inf_X(f) = -\infty$.
- In any other cases, $\inf_X(f)$ is the largest lower bound of f over X. We have

$$\inf_X(f) = \max\{z, \text{such that } z \leq f(x) \quad \forall x \in X\}$$

Definition 1.8.2 — Minimum of a function. Let X a subset of a Banach space and $f : X \mapsto \mathbb{R}$.

- If there exists $x^\star \in X$ such that $f(x^\star) = \inf_X f$, then
  1. We say that the infimum of f over X is **attained**.
  2. We call this infimum the **minimum**.

3. We call $x^\star$ a **minimizer** of f over X

- The set of minimizers is denoted arg min, by definition :

$$z \in \arg\min_{x\in X} f(x) \iff \left(f(z) = \inf_{x\in X} f(x) \text{ and } z \in X \right).$$

Note that the infimum is not attained iff $\arg\min_{x\in X} f(x) = \emptyset$

By convention, we have $\arg\min_{x\in X} f(x) = \emptyset$ if $f(X)$ does not have a minimum. Hence, arg min always exists even if min does not always exists.

Exercice 1.5

Give the minimizers, if they exist of the problem $\inf_{x\in X} f(X)$, where

1. $X = \mathbb{R}$ and $f(x) = x^2 + 1$.
2. $X = \mathbb{R}$ and $f(x) = \cos(x)$.
3. $X = \mathbb{R}$ and $f(x) = e^x$.
4. $X = [1,2]$ and $f(x) = x^2 + \pi$.
5. $X =]1,2]$ and $f(x) = x^2 + \pi$.

Solution to Exercice 1.5

We have

1. $\arg\min_{x\in X} f(x) = \{0\}$
2. $\arg\min_{x\in X} f(x) = \{(2k+1)\pi, k \in \mathbb{Z}\}$
3. $\arg\min_{x\in X} f(x) = \emptyset$
4. $\arg\min_{x\in X} f(x) = \{1\}$
5. $\arg\min_{x\in X} f(x) = \emptyset$

Definition 1.8.3 — Local minimum. Let X be a subset of a Banach space. We say that $x^\star$ is a **local minimizer** of f over X if and only if there exists $r > 0$ such that if B is the ball of radius r and of center $x^\star$, then $x^\star$ is a minimizer of f on $X \cap B$. In this case the value $f(x^\star)$ is called a **local minimum**.

If there is a risk of confusion, we sometimes say that the minimizer of f over X are **global minimizers** (as opposed to "local minimizer") and we often say that the minimum is the **global minimum**. Hence a "global minimum" is always a "local minimum" and a "global minimizer" is always a "local minimizer". The converse is false. Note also that there might exists more than one local minimum. The plural of "minimum" is **minima** (in french also, it is a latin word) and not "minimums".

1.9 The effect of convexity on minimization problems

1.9.1 Globalization of minima

Theorem 1.9.1 — no local minima.

1. If f is convex over the convex set X, then local minima of f over X are global minima.
2. If f is a C^1 and convex function over the convex set X, then if x is a critical point of f (a point where $\nabla f(x) = 0$), then x is a global minimum of f over X.

Proof

1. We prove the first item by *reductio ad absurdum*. Let $x_0 \in X$ be a local minimizer of f over X, i.e. there exists $r > 0$ such that:

$$\forall y \in X \text{ such that } \|y - x_0\| < r \text{ then } f(y) \geq f(x_0). \tag{1.2}$$

Suppose that x_0 is not a global minimum of f over X, i.e. there exists $x_1 \in X$ such that

$$f(x_1) < f(x_0). \tag{1.3}$$

For any $\theta \in [0,1]$, we introduce $x_\theta = \theta x_1 + (1-\theta)x_0$. The point x_θ belongs to the convex set X for any $\theta \in [0,1]$. Moreover, we have, if $\theta \neq 0$

$$f(x_\theta) = f(\theta x_1 + (1-\theta)x_0) \leq \theta f(x_1) + (1-\theta) f(x_0) < f(x_0)$$

If θ is close to 0, we can manage to have $\|x_\theta - x_0\| < r$. This contredicts (1.2).

2. If x is a critical point, then $\nabla f(x) = 0$. Because f is convex, we use Theorem 1.6.1 to obtain, for any y :

$$f(y) \geq f(x) + \langle \underbrace{\nabla f(x)}_{=0}, y - x\rangle = f(x)$$

Hence x is a global minimum.

1.9.2 Strict convexity

Theorem 1.9.2 — Strict convexity and uniqueness. If f is strictly convex on the convex set X, then f admits at most one global minima (and hence at most one local minima) on X.

Proof

Let $x_1 \neq x_2$ be two elements of X which are global minimizers of f. By convexity of X, it holds that $\frac{x_1+x_2}{2} \in X$ and by strict convexity of f we have,

$$\begin{aligned} f\left(\frac{x_1+x_2}{2}\right) &< \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2) \\ &< \frac{1}{2} \min_{y \in X} f(y) + \frac{1}{2} \min_{y \in X} f(y) = \min_{y \in X} f(y), \end{aligned}$$

which is impossible.

CHAPTER 2

Existence

In this chapter, we review the theorem of existence of minimisation problems seen in the previous year. We suppose that the reader is familiar with the following theorem, coined as the Weierstrass Theorem :

> If f is a continuous function on $X \subset \mathbb{R}^n$, where X is a non-empty bounded and closed set, then the function f admits minmizers and maximizers.

The above theorem is not to be confused with the following result, known as the Bolzano-Weierstrass Theorem :

> Any bounded sequence of $\mathbb{R}^n$ admits a convergent subsequence.

Note that Bolzano-Weierstrass Theorem is key to proving Weierstrass Theorem (see for instance Theorem 2.2.2 which is a generalization of Weierstrass and its proof), hence the confusion. The goal of this chapter is to improve Weierstrass Theorem in several directions, notably by studying the infinite dimensional case. The key idea is that, in infinite dimension, the Boltzmann-Weierstrass Theorem is not true anymore and one usually resorts to the Banach-Alaoglu Theorem. One of the goal of this chapter is to introduce these notions and guide the reader through the main ideas of infinite dimensional optimization. In this chapter, we will suppose that the reader is familiar with the notion of a Banach space. This notion is briefly recalled in the definition below

> Let X be a metric vector space (a vector space with a norm). We say that the sequence $(u_n)_n$ is a Cauchy sequence iff
>
> $$\lim_{n\to+\infty}\left(\sup_{k\geq n, k'\geq n}\|u_k - u_{k'}\|\right) = 0.$$
>
> If every Cauchy sequence converges, X is said to be complete. A complete metric vector space is called a **Banach** space.

2.1 Minimizing sequences

The proof of existence of minimization problems relies strongly on three notions : the existence of minimizing sequences, lower semi-continuity and compactness. We study the first notion in the following proposition.

Proposition 2.1.1 — Minimizing sequence. Let X be any kind of set and $f : X \mapsto \mathbb{R}$. We say that $(x_n)_n$ is a minimizing sequence for the problem $\inf_X f$ if $x_n \in X$ for every n and

$$\lim_{n\to+\infty} f(x_n) = \inf_X f.$$

If X is not empty, there always exists a minimizing sequence.

It is very important to remember the hypothesis of any given proposition. In the above proposition, there is none. It means that the concept of minimizing sequence is very important because there **always** exists a minimizing sequence, no matter what.

Proof

We recall that $\inf_X f$ is the largest lower bound of $f(X)$, that is the largest $m \in \mathbb{R}$ such that $m \le f(x)$ for each $x \in X$. By definition $\inf_X f = +\infty$ if X is empty and $\inf_X f = -\infty$ if $f(X)$ has no lower bound.

1. Suppose that $\inf_X f = -\infty$. Hence there is no lower bound of f over X, that is for every $n \in \mathbb{N}$, $-n$ is not a lower bound, hence $\exists x_n \in X$ such that $f(x_n) < -n$, and then $(x_n)_n$ is a minimizing sequence.
2. Suppose now that $\inf_X f = m \in \mathbb{R}$. Fix $n \in \mathbb{N}$
   (a) Because $m + \frac{1}{n}$ is not a lower bound of f over X, then there exists $x_n \in X$ such that $f(x_n) < m + \frac{1}{n}$
   (b) Take the x_n defined above, because m is a lower bound of f over X, we always have $f(x_n) \ge m$

   Hence the sequence $(f(x_n))_n$ converges to m and $(x_n)_n$ is a minimizing sequence.

The second important notion is the one of **lower semi-continuity**. It bears a barbaric name, but is just an extension of the notion of continuity. First, we must remember what is continuity. If f is a continuous function at a certain point x, when we study the values of $f(x_n)$ as x_n approaches x, we find that $f(x_n)$ tends towards $f(x)$. In other words, f is continuous if "f of the limit" is **equal** to "the limits of the f's". A function is lower semi-continuous if "f of the limit" is **smaller** than "the limits of the f's".

Definition 2.1.1 — Lower semi-continuity. Let $f : X \mapsto \mathbb{R}$ where X is a Banach space. The function f is l.s.c at $x_0 \in X$ iff for every sequence $x_n \in X$ that converges to x_0, we have :

$$\liminf_{n\to+\infty} f(x_n) \ge f(x_0).$$

Where

$$\liminf_{n\to+\infty} u_n \overset{\text{def}}{=} \lim_{n\to+\infty} \inf_{k\ge n} u_k = \sup_n \inf_{k\ge n} u_k$$

Here the notion of lim inf seems a little bit complicated, but it is a life-saver for mathematicians. Indeed consider any sequence of real number $(u_n)_n$, in general we cannot talk about $\lim\limits_{n\to+\infty} u_n$ because we do not know if it exists. But consider the sequence $(v_n)_n$ given by $v_n = \inf_{k\ge n} u_k$, then $(v_n)_n$ is an increasing sequence so that it always admits a limit. Possibly this limit is equal to $+\infty$ or $-\infty$, but it exists. Similarly introducing $w_n = \sup_{k\geq n} u_k$, then w_n is a decreasing sequence. Now define

$$\liminf_{n\to+\infty} u_n = \lim_{n\to+\infty} v_n \text{ and } \limsup_{n\to+\infty} u_n \overset{\text{def}}{=} \lim_{n\to+\infty} w_n.$$

Then both $\liminf u_n$ and $\limsup u_n$ are objects that **always** exists. Moreover we have $v_n \leq u_n \leq w_n$ for each n, this means that if $(v_n)_n$ converges to the same limit as $(w_n)_n$ then the whole sequence $(u_n)_n$ converges. Hence we established the theorem

$$\left(\liminf_{n\to+\infty} u_n = \limsup_{n\to+\infty} u_n\right) \Rightarrow \lim_{n\to+\infty} u_n \text{ exists.}$$

Note that this theorem uses objects such as $\liminf$ or $\limsup$ that always exists to conclude on the existence of an object that might not exist (and mathematicians HATE objects that might not exist). It is not hard to prove that the above implication is indeed an equivalence, so that the existence of the limit of sequence of real numbers implies that the $\liminf$ and $\limsup$ must coïncide. Anyway, if one is too afraid to use $\liminf$[1], one could use the following equivalent definition of lower semi-continuity :

Proposition 2.1.2 — Equivalent definitions. Let $f : X \mapsto \mathbb{R}$ where X is a Banach space. The function f is l.s.c at $x_0 \in X$ iff for every sequence $x_n \in X$ that converges to x_0, such that $\lim_{n\to+\infty} f(x_n)$ exists (possibly equal to $+\infty$ or $-\infty$), then

$$\lim_{n\to+\infty} f(x_n) \geq f(x_0).$$

In order to understand the notion of lower semi-continuity, we propose an exercise

Exercice 2.1

Let $a, b, c \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = \begin{cases} a \text{ if } x < 0 \\ b \text{ if } x = 0 \\ c \text{ if } x > 0 \end{cases}$. Show that

1. f is l.s.c. iff $b \leq a$ and $b \leq c$,
2. f is u.s.c. iff $b \geq a$ and $b \geq c$,
3. f is continuous iff $a = b = c$.

2.2 Existence of minimum in finite dimension

In this section, we give sufficient conditions that ensure existence of a global minimum. Before dwelling into the subject, we study some cases where there is no global minima. The three examples below are characteristic of three different phenomenon that the hypotheses will prevent .

Proposition 2.2.1 The three following problems do not admit a global minimum :

- No l.s.c

$$\inf_{x\in\mathbb{R}} g(x) = 0 \text{ if } g(x) = \begin{cases} x^2 \text{ when } x \neq 0 \\ 1 \text{ when } x = 0 \end{cases}$$

[1] and one would be a fool, because the sole purpose of $\liminf$ is to be helpfull to mathematicians

- **No closedness**

$$\inf_{x\in]1,2]} x^2 = 1$$

- **Infimum at infinity**

$$\inf_{x\in \mathbb{R}} e^{-x} = 0$$

These counterexamples are sharp in the sense that if we manage to counter the three phenomenon at play, we can ensure the existence of a minimum. The main theorem of existence is as follows:

Theorem 2.2.2 — Existence of minimum (finite dimension). The function f admits a minimum on $X \subset \mathbb{R}^d$ if

- **l.s.c :** The function f is lower semi-continuous on X.
- **closedness:** The set X is non-empty and closed.
- **coercivity:** For every sequence $(x_n)_n \in X$, we have

$$\text{If } \lim_{n\mapsto +\infty} \|x_n\| = +\infty \quad \text{then} \quad \lim_{n\mapsto +\infty} f(x_n) = +\infty$$

Proof

1. By Proposition 2.1.1, because X is non-empty, there exists a minimizing sequence $(x_n)_n \in X$ such that $\lim_{n\to+\infty} f(x_n) = \inf_X f$.
2. This sequence $(x_n)_n$ is bounded. Indeed suppose it is not the case, then, up to a subsequence which we do not relabel, we have :

$$\lim_{n\mapsto +\infty} \|x_n\| = +\infty.$$

Then the following equalities lead to a contradiction:

$$\inf_X f = \lim_{n\mapsto+\infty} f(x_n) \underbrace{=}_{\text{coercivity}} +\infty.$$

3. Because $(x_n)_n$ is bounded, then $(x_n)_n$ has a convergent subsequence. Up to relabeling, we suppose that $(x_n)_n$ converges. We denote $x^\star$ the limit of $(x_n)_n$
4. X is closed so that $x^\star \in X$.
5. f is lower semi-continuous so that

$$\inf_X f = \lim_{n\mapsto+\infty} f(x_n) \underbrace{\geq}_{\text{l.s.c}} f(\lim_{n\mapsto+\infty} x_n) = f(x^\star)$$

6. Hence $x^\star$ is a minimizer of f over X.

Before giving some examples, we focus on proving the different hypothesis (continuity, closedness, coercivity). The continuity hypothesis is easy to prove. The closedness is easy to prove, when X is in the so-called standard form. We say that X is in standard form when it is defined by inequalities and/or equalities.

Definition 2.2.1 — Standard form. Let $X \subset \mathcal{B}$, where $\mathcal{B}$ is a Banach space. We say that X is in standard form if there exists a finite number of functions $g_i : \mathcal{B} \mapsto \mathbb{R}$

such that

$$X = \{x \text{ such that } g_i(x) \leq 0 \text{ for every } i \in \mathcal{I} \text{ and } g_i(x) = 0 \text{ for every } i \in \mathcal{E}\},$$

where $\mathcal{I}$ and $\mathcal{E}$ are finite sets of indices.

Note that $g_i(x) = 0$ is equivalent to $g_i(x) \leq 0$ and $-g_i(x) \leq 0$. Hence if a set X is in standard form, then it can be described by inequalities only, as

$$X = \{x \text{ such that } g_i(x) \leq 0 \text{ for every } i \text{ and } - g_i(x) \leq 0 \text{ for every } i \in \mathcal{E}\}.$$

In every theorem, we will discuss the behavior of the theorem when X is re-defined with the above method and is described by inequalities only. The above transformation is coined **transforming equalities into inequalities**.

Proposition 2.2.3 — Closedness of X when in standard form. Suppose that X is a subset of a Banach space and that X is in a standard form,

$$X = \{x \text{ such that } g_i(x) \leq 0 \text{ for every } i \in \mathcal{I} \text{ and } g_i(x) = 0 \text{ for every } i \in \mathcal{E}\},$$

If for every $i \in \mathcal{I}$, the function g_i is l.s.c, and if for every $i \in \mathcal{E}$, the function g_i is continuous, then X is closed.

Note that if the equalities are transformed into inequalities in the description of X, the hypothesis transforms into : for every $i \in \mathcal{E}$, both g_i and $-g_i$ must be l.s.c. This is equivalent to asking for continuity of g_i.

Proof

Take a sequence $(x_n)_n$ of elements of X that converges to some $x^\star$.

1. For any $i \in \mathcal{E}$, by continuty of g_i we have $g_i(x_n) = 0$ for each n implies that $g_i(x^\star) = 0$.
2. For any $i \in \mathcal{I}$, we have $g_i(x_n) \leq 0$, hence $v_n = \inf_{k \geq n} g(x_k) \leq 0$ for every n and

$$\liminf g_i(x_n) = \sup_n v_n \leq 0.$$

And by l.s.c. of g_i, we have $g_i(x^\star) \leq 0$.

Hence $x^\star \in X$ and X is closed.

We turn ourselves to proving the coercivity assumption. There is usually two different ways to prove this assumption.

Proposition 2.2.4 — Ensuring coercivity. Let X be a subset of a Banach space. The coercivity assumption holds if either one of the following assumption is true

- The set X is bounded.
- There exists a function $r : \mathbb{R}^+ \to \mathbb{R}$ such that $\lim_{t \mapsto +\infty} r(t) = +\infty$ and

$$\forall x \in X, \quad f(x) \geq r(\|x\|)$$

Exercice 2.2

Show that the following problems admit a solution

1. $\min\limits_{x^2+y^2\leq 1} 3x+5y$
2. $\min\limits_{x\geq 0, x^2+y^2\leq 1} 3x+5y$
3. $\min\limits_{(x,y)\in\mathbb{R}^2} 3x^4+2y^4-10x^3-5xy^2$

Solution to Exercice 2.2

1. Here $f(x,y)=3x+5y$ and $g_1(x,y)=x^2+y^2-1$. The functions f and g_1 are continuous and X is bounded.
2. Here $f(x,y)=3x+5y$ and $g_1(x,y)=x^2+y^2-1$ and $g_2(x,y)=-x$. We have two inequality constraints here. The functions f, g_1 and g_2 are continuous and X is bounded (because of g_1).
3. Here $f(x,y)=3x^4+2y^4-10x^3-5xy^2$. The function f is continuous. Since there is no constraints, the set X is closed. Indeed $\mathbb{R}^2$ is a closed set. We use the ∞ norm to prove coercivity, denoting $M=(x,y)$, we have

$$x^4+y^4\geq \|M\|_\infty^4$$

and then

$$\begin{array}{rcl} f(M) & \geq & 2\|M\|_\infty^4-10\|M\|_\infty^3-5\|M\|_\infty^3 \\ & \geq & g(\|M\|_\infty) \text{ avec } g(t)=2t^4-15t^3. \end{array}$$

Because $\lim\limits_{t\to+\infty} g(t)=+\infty$, then f is coercive.

Exercice 2.3

In the following examples, explain why the theorems don't apply. If applicable, prove that there is no minimizer

1. $\inf\limits_{x>0} 3x$
2. $\inf\limits_{(x,y)\in\mathbb{R}^2} x^2+y^2+100xy$

Solution to Exercice 2.3

1. Here the set $X=\mathbb{R}^{+*}$ is not closed. The infimum is 0 (because 0 is a lower-bound and there is no lower-bound greater than 0). And there is no $x\in X$ such that $3x=0$. Hence there is no miminizer to this problem.
2. Here there is no coercivity. Indeed $f(n,n)=-98n^2\mapsto-\infty$ when $n\mapsto+\infty$.

2.3 Existence of minimum in the infinite dimension case

The attentive reader will have noticed that Proposition 2.2.3 and 2.2.4 are stated for X a subset of a Banach space. But Theorem 2.2.2 that proves the existence of a minimizer requires X to be a subset of a finite dimensional space. In infinite dimension, the bounded closed sets are not necessarily compact, meaning that there might exists a bounded sequence that does not admit a limit, even up to a subsequence. The archetypal example is given by the exercise below :

Exercice 2.4

Let $\ell^2(\mathbb{R})$ be the space of sequences $(u[i])_{i\in\mathbb{N}}$ of real numbers such that

$$\sum_{i=0}^{+\infty}(u[i])^2 < +\infty.$$

We can prove that $\ell^2(\mathbb{R})$ is a Banach space, it is even a Hilbert space when endowed with the scalar product

$$\langle u, v\rangle = \sum_{i=0}^{+\infty} u[i]v[i].$$

Consider the sequence $(u_n)_n$ where for each n, $u_n \in \ell^2(\mathbb{R})$ is defined by

$$u_n[i] = 0 \text{ for all } i \neq n \text{ and } u_n[n] = 1$$

1. Show that for each n, $\|u_n\|_{\ell^2(\mathbb{R})} = 1$
2. Assume that there exist a subsequence of $(u_n)_n$ that converges to some $u \in \ell^2(\mathbb{R})$. Show that $u = 0$
3. Show that no subsequence of $(u_n)_n$ can converge to 0.
4. Show however that for every $v \in \ell^2(\mathbb{R})$, then $\langle u_n, v\rangle$ converges to 0 as n goes to $+\infty$.

Solution to Exercice 2.4

1. We compute $\|u_n\|_{\ell^2} = \sqrt{\sum_{i=0}^{+\infty} u_n[i]^2} = \sqrt{u_n[n]^2} = 1$
2. Call the subsequence $(u_n)_n$. If $(u_n)_n$ converges to u then $\langle u_n, u\rangle$ converges to $\|u\|^2$. But
$$\langle u_n, u\rangle = \sum_{i=0}^{+\infty} u_n[i]u[i] = u[n].$$
Because $u \in \ell^2$, then $u[n]$ must converge to 0 as n goes to $+\infty$. Therefore $\|u\|^2 = 0$ and $u = 0$
3. Suppose that there exists a subsequence of $(u_n)_n$, still denoted $(u_n)_n$ that converges towards 0. Then $\|u_n - 0\|$ must converge to 0, which is impossible because $\|u_n\| = 1$ for every n.
4. Let $v \in \ell^2(\mathbb{R})$, then $\langle u_n, v\rangle = v[n]$ and $v[n]$ converges to 0 as n goes to infinity because the series $\sum_i v[i]^2$ converges.

The above example shows that the proof of Theorem 2.2.2 might fail to extend to infinite dimensional problems, indeed coercivity ensure that minimizing sequences are bounded but bounded sequences do not necessary converge (even up to a subsequence). The future of optimization in infinite dimension looks very grim, until one of the most powerfull theorem of mathematics pops out to save the day, it is the celebrated **Banach-Alaoglu** theorem. It is a very general theorem which is not easy to state in its full glory, we will restrict ourselves to a version in a Hilbert space (there exists versions in Banach spaces, but we do not really care here).

Theorem 2.3.1 — Hilbert version of Banach-Alaoglu. Let E be a Hilbert space, we say that a sequence $(u_n)_n$ of vectors of E weakly-converges to u and we denote $u_n \rightharpoonup u$

iff

$$\forall v \in E, \lim_{n \to +\infty} \langle u_n, v\rangle = \langle u, v\rangle.$$

Any bounded sequence of elements of E admits a subsequence that weakly-converges.

This is pure magic: Indeed Weierstrass Theorem is true in infinite dimension if we just replace the notion of convergence by the notion of weak-convergence. We just have to change the name "Weierstrass" by "Banach-Alaoglu". The proof of the above theorem is **really** difficult. We will nevertheless try to show some elements of proof if we suppose that the Hilbert space E is separable. An Hilbert space is said to be separable if there exists an orthonormal family $(e_i)_{i\in\mathbb{N}}$ which is denombrable such that for each $u \in E$, we have

$$u = \sum_{i=0}^{+\infty} u[i]e_i \text{ with } u[i] = \langle u, e_i\rangle.$$

In words, a Hilbert space is separable if and only if it admits a basis which is denombrable. It turns out that most of the Hilbert spaces are separable and a popular quote amongst mathematicians is : "I do not know what a non-separable Hilbert space is. And if it exists, I do not want to know about it". Coming back to the proof of Banach-Alaoglu, suppose E is separable and take $(u_n)_n$, a bounded sequence of elements of E. For each i the sequence of real numbers $(u_n[i])_n$ is bounded and converges up to a subsequence. Successively extract subsequences (it is called a **diagonal extraction argument**) then for each i, $(u_n[i])_n$ converges to some $u[i]$. It is not difficult to prove then that weak convergence is equivalent to the convergence of coordinates.

It is a simple exercise to follow the proof of Theorem 2.2.2 and to obtain the following

Exercice 2.5: Existence of minimum

Let E be a Hilbert space and $X \subset E$. We say that

- a set X is weakly-closed if for each sequence $(x_n)_n \in X$ that weakly converges to $x \in E$, then $x \in X$.
- a function f is weakly-lower semi-continuous iff for each $(x_n)_n \in X$ that weakly converges to $x \in X$ then $\liminf f(x_n) \geq f(x)$.

If

- **l.s.c.:** The function f is weakly-lower-semi-continuous on X .
- **closedness:** The set X is non-empty and weakly-closed.
- **coercivity:** For every sequence $(x_n)_n \in X$, we have

$$\text{If } \lim_{n\mapsto+\infty} \|x_n\| = +\infty \quad \text{then} \quad \lim_{n\mapsto+\infty} f(x_n) = +\infty$$

Solution to Exercice 2.5

Trivial !! (yes it is), replace Weierstrass theorem by Banach-Alaoglu

The above theorem is kind of useless (by the way it is not a theorem, it is an exercise). The reason is that it is difficult to show that a function is weakly-lower-semi-continuous or that a set is weakly-closed. Except in a very particular case, when the function f is convex and the set X is convex. Indeed in a Hilbert space, every closed convex set is weakly closed and every lower-semi-continuous function f is weakly-lower-semi-continuous. Hence we have the following theorem :

Theorem 2.3.2 — Existence in infinite dimension, convex case. Let E be a Hilbert space and $X \subset E$. If X is convex and f is convex on X, there exists a minimum to f over X if

- **l.s.c.:** The function f is lower-semi-continuous on X .
- **closedness:** The set X is non-empty and closed.
- **coercivity:** For every sequence $(x_n)_n \in X$, we have

$$\text{If } \lim_{n \mapsto +\infty} \|x_n\| = +\infty \quad \text{then} \quad \lim_{n \mapsto +\infty} f(x_n) = +\infty$$

We finish this section by a very important example that does not require the Banach-Alaoglu theorem and all the machinery developed above. The following theorem is important because it allows to prove Hahn-Banach Theorem (see Proposition 3.2.2 later). The fun part is that if Banach-Alaoglu is the most important theorem of analysis, Hahn-Banach is the second !! So kudos to Banach[2], and lets prove : **The projection Theorem**.

Theorem 2.3.3 — Projection on a convex set. Let E be a Hilbert space and suppose that $\mathcal{C} \subset E$ is a closed convex set. For every $x \in E$, there exists a unique solution to

$$\min_{y \in \mathcal{C}} \frac{1}{2}\|y - x\|^2.$$

This solution is denoted $\pi_{\mathcal{C}}(x)$ and it verifies

$$\langle \pi_{\mathcal{C}}(x) - x, y - \pi_{\mathcal{C}}(x) \rangle \geq 0 \quad \forall y \in \mathcal{C}. \tag{2.1}$$

If in addition $\mathcal{C}$ is a closed vector space, we have $\langle \pi_{\mathcal{C}}(x) - x, v\rangle = 0$ for each $v \in \mathcal{C}$.

The main idea of the proof is that the minimizing sequence is a Cauchy-sequence and hence converges. The reason is that the function we minimize is the norm and it gives a lot of information about minimizing sequences.

Proof

As a preliminary, remember the parallellogram identity valid for any $c, d \in E$

$$\|c\|^2 + \|d\|^2 = \frac{1}{2}\left(\|c + d\|^2 + \|c - d\|^2\right)$$

This shows that for every vectors x, a, b in E, we have

$$\|x - \frac{a+b}{2}\|^2 + \|\frac{a-b}{2}\|^2 = \frac{1}{2}\left(\|x - b\|^2 + \|x - a\|^2\right). \tag{2.2}$$

Let $(y_n)_n$ be a minimizing sequence and denote $d_n = \|y_n - x\|^2$. Denote $d = \inf_{y \in \mathcal{C}} \frac{1}{2}\|y - x\|^2$. We have, by (2.2)

$$\|x - \frac{y_n + y_p}{2}\|^2 + \|\frac{y_n - y_p}{2}\|^2 = d_n + d_p.$$

[2] When learning modern analysis, it is common for young mathematicians to consider Stefan Banach as a half-god for the discovery of Hahn-Banach and Banach-Alaoglu theorem. Then they learn about Banach-Tarski theorem and understand that Stefan Banach multiplies balls faster than Jesus Christ can multiply fish and bread. And the young mathematician will certainly hold Banach for the most influential analyst of all time. At least until he hears about a certain Terence Tao...

By convexity of $\mathcal{C}$, $\frac{y_n+y_p}{2} \in \mathcal{C}$, so that $\|x - \frac{y_n+y_p}{2}\|^2 \geq 2d$. We then have

$$\|y_n - y_p\|^2 \leq 4(d_n + d_p - 2d).$$

Using that $(d_n)_n$ converges to d, this proves that the sequence $(y_n)_n$ is a Cauchy sequence. Because E is complete, then $(y_n)_n$ converges to some $\pi_{\mathcal{C}}(x)$. Because $\mathcal{C}$ is closed, then $\pi_{\mathcal{C}}(x) \in \mathcal{C}$. The function $y \mapsto \|y-x\|^2$ is continuous, hence $\|\pi_{\mathcal{C}}(x) - x\|^2 = d$ and we have a solution to the minimisation problem. Uniqueness stems from the fact that $\mathcal{C}$ is convex and strict convexity of $y \mapsto \|y-x\|^2$.
Now, $\forall y \in \mathcal{C}$, $d = y - \pi_{\mathcal{C}}(x)$ is an admissible direction at point $\pi_{\mathcal{C}}(x)$, in the sense that for all $\theta \in [0,1]$, $\pi_{\mathcal{C}}(x) + \theta d \in \mathcal{C}$, by convexity of $\mathcal{C}$. We write the (exact) Taylor expansion of $f(y) = \frac{1}{2}\|y-x\|^2$ around the point $\pi_{\mathcal{C}}(x)$ to find

$$f(\pi_{\mathcal{C}}(x) + \theta d) = f(\pi_{\mathcal{C}}(x)) + \theta\langle d, \pi_{\mathcal{C}}(x) - x\rangle + \frac{\theta^2}{2}\|d\|^2.$$

By letting θ small, it is impossible to have $\langle d, \pi_{\mathcal{C}}(x) - x\rangle < 0$, hence (2.1). Now of $\mathcal{C}$ is a closed vector space, take $y = \pi_{\mathcal{C}}(x) \pm v$, then $y \in \mathcal{C}$ and plug these y in (2.1) to obtain $\langle \pi_{\mathcal{C}}(x) - x, v\rangle = 0$.

Chapter 3

First order conditions

3.1 Presentation of the main result

3.1.1 Introduction

We suppose that the reader is familiar with the following proposition

Proposition 3.1.1 — Euler first order conditions (open set). Let E be a Hilbert space and $X \subset E$ be an **open** set. Let f be a C^1 function on X. If $x^\star$ is a local minimum of f on X, then $\nabla f(x^\star) = 0$

Proof

Since $x^\star \in X$ and X is open, then for any choice of direction $d \in E$, then $x^\star + \varepsilon d \in X$ for small ε. Moreover, for small ε, we also know that we have $f(x^\star + \varepsilon d) \geq f(x^\star)$. Performing a first order Taylor exansion, we obtain

$$0 \leq f(x^\star + \varepsilon d) - f(x^\star) = \varepsilon \langle \nabla f(x^\star), d\rangle + o(\varepsilon)\,.$$

This proves that it is impossible to have $\langle \nabla f(x^\star), d\rangle < 0$. Specializing d to the choice $d = -\nabla f(x^\star)$, we obtain $\nabla f(x^\star) = 0$.

The above theorem is coined as **Euler first order condition**. In view of Theorem 2.2.2, in order to ensure existence, we expect the set X to be a closed set. Hence the above theorem is only usefull if X is both closed and open, that is if $X = E$. This is the reason why the above theorem is often calle the **unconstrained first order condition**. The goal of this chapter is to state the equivalent of Euler's first order condition when X is not an open set, but a closed set defined by equalities and inequalities. For that purpose, we recall the notation of the so-called **standard form**

Let E be a Hilbert space and $X \subset E$. We suppose that there exists a **finite** number of functions $(g_i)_i$ and finite sets of indices $\mathcal{I}$ and $\mathcal{E}$ such that

$$X = \{x \in E \text{ such that } g_i(x) \leq 0, \quad \forall i \in \mathcal{I} \text{ and } g_i(x) = 0, \quad \forall i \in \mathcal{E}\}\,.$$

Such an X is then said to be in **standard form**.

In this chapter, we begin by stating the result in Section 3.1.2, it is Theorem 3.1.4. We then prove some premilinary lemmas in Section 3.2, we introduce some new concepts in Section 3.3, notably the notion of qualification of constraints. Finally, in Section 3.4.1, we prove Theorem 3.1.4.

3.1.2 Statement of KKT Theorem

Definition 3.1.1 — Active constraints. For any $x \in X$, denote $\mathcal{A}_x$ the set of indices $i \in \mathcal{I}$ such that $g_i(x) = 0$. This set is called the set of active constraints at x.

$$\mathcal{A}_x = \{i \in \mathcal{I} \text{ such that } g_i(x) = 0\}\,.$$

We say that an inequality constraint g_i is active at x if and only if $g_i(x) = 0$

There exists a technical condition for KKT Theorem. This condition is called the qualification of constraints. We do not want to discuss about constraint qualification here, we just give two different conditions that both lead to constraint qualification :

Proposition 3.1.2 — LICQ condition. Let $x \in X$, if the family $(\nabla g_i(x))_{i \in \mathcal{E} \cup \mathcal{A}_x}$ is linearly independent, we say that the linear independence constraint qualification (LICQ) holds at x. If (LICQ) holds at x, then the constraints are qualified at x.

In the convex setting, there is another simpler proposition that implies qualification of constraints.

Proposition 3.1.3 — Slater's condition. Suppose that for each $i \in \mathcal{I}$, g_i is convex and forall $i \in \mathcal{E}$, g_i is affine. Then X is convex. Moreover if there exists a point $x \in X$ such that for each i

$$g_i(x) < 0 \text{ or } g_i \text{ is affine},$$

then the constraints are qualified everywhere in X.

We are now ready to state the main result of this chapter.

Theorem 3.1.4 — KKT. Let E be a Hilbert space and $X \subset E$ be a set in standard form. Suppose that f and $(g_i)_{i \in \mathcal{I} \cup \mathcal{E}}$ are a C^1 function on X. Denote $\mathcal{L}$ the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i \in \mathcal{I} \cup \mathcal{E}} \lambda_i g_i(x), \text{ with } \lambda_i \in \mathbb{R}\ \forall i \text{ and } \lambda_i \geq 0\ \forall i \in \mathcal{I}$$

If $x^\star$ is a local minimum of f on X and the constraints are qualified at point $x^\star$, there exists $\lambda^\star$ such that

$$\begin{cases} \nabla_x \mathcal{L}(x^\star, \lambda^\star) = 0 & \\ x^\star \in X \text{ and } \lambda_i^\star \geq 0 & \text{forall } i \in \mathcal{I} \\ \lambda_i^\star g_i(x^\star) = 0 & \text{forall } i \end{cases}$$

What is the difference between the above theorem (known as KKT Theorem) and Euler's first order conditions ? First, there is an extra hypothesis, the qualification of constraints. This hypothesis might seem confusing, but it is rather technical and we are purposely vague about it for the time being. Second, the unknowns are $(x^\star, \lambda^\star)$ and $\lambda^\star$

are additional unknowns that we have to find. We recognize that the condition $\nabla f = 0$ turns into $\nabla \mathcal{L} = 0$, this give us as many equations as number of unknown in $x^\star$. Finally, we have an additional set of equations which is $\lambda_i^\star g_i(x^\star) = 0$ which are known as **the complementary conditions**. We see that there are as many complementary conditions as the number of constraints (and hence as the size of $\lambda^\star$).

3.1.3 Application of KKT theorem

Before proving KKT theorem, we give two applications of the theorem

Exercice 3.1

Find every rectangle in $\mathbb{R}^2$ of maximal surface with perimeter smaller than $P \in \mathbb{R}^{+*}$.

Solution to Exercice 3.1

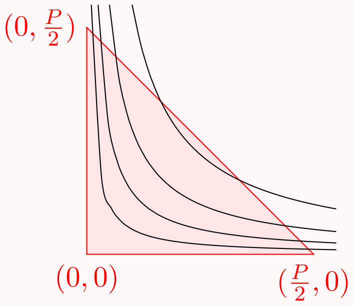
We split the solution into several different simple steps.

- **Step -1 : Standard form** Let x_1 and x_2 be the lengths of the rectangle, let $x = (x_1, x_2)$ and denote

$$f(x) = -x_1 x_2 \text{ and } g(x) = \begin{pmatrix} 2(x_1 + x_2) - P \\ -x_1 \\ -x_2 \end{pmatrix}.$$

  If $X = \{x \text{ such that } g_i(x) \leq 0 \text{ forall } i\}$ then we aim at minimizing f over X.

- **Step 0 : Explanatory figure** The set X is displayed in red and some level-sets of f are displayed in black.



- **Step 1 : Existence of a minimum** The function f is continuous and X is closed (by continuity of g). For each $x \in X$, we have $\|x\|_1 \leq \frac{P}{2}$, and therefore X is bounded. There exists a minimum and a maximum of f on X.
- **Step 2 : Qualification of constraints** The constraints are affine hence qualified by Slater's condtion
- **Step 3 : Solving KKT** The Lagrangian is given by

$$\mathcal{L}(x, \lambda) = -x_1 x_2 + \lambda_1(2x_1 + 2x_2 - P) - \lambda_2 x_1 - \lambda_3 x_2$$

  We write down KKT equations : Find x and λ such that $g_i(x) \leq 0$ and $\lambda_i \geq 0$ for $i \in \{1, 2, 3\}$ and such that :

$$\nabla \mathcal{L}(x, \lambda) = 0 \text{ and } \lambda_i g_i(x) = 0 \quad \forall i$$

  The first equation is

$$\begin{pmatrix} -x_1 \\ -x_2 \end{pmatrix} + \lambda_1 \begin{pmatrix} 2 \\ 2 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \lambda_3 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = 0.$$

  - If $\lambda_1 = 0$, then $x_1 = -\lambda_3$ and $x_2 = -\lambda_2$. From $\lambda_3 g_3(x) = 0$ we deduce that either λ_3 or x_2 is equal to 0, that is either x_1 or x_2 is zero. In any case $f = 0$.
  - If $\lambda_1 \neq 0$ then we have $g_1(x) = 0$, hence $x_1 + x_2 = \frac{P}{2}$. If $\lambda_2 \neq 0$ (resp $\lambda_3 \neq 0$), then $0 = g_2(x) = -x_1$ (resp. $0 = x_2$) and $f = 0$. If $\lambda_2 = \lambda_3 = 0$ we have $x_1 = x_2 = \frac{\lambda_1}{2}$ and from $x_1 + x_2 = \frac{P}{2}$, we deduce that $x_1 = \frac{P}{4}$ and $f = -\frac{P^2}{16}$.
- **Step 4 : Conclusion** For each possible KKT point, the value of f is either 0 or $-\frac{P^2}{16}$. There exists a minimum which is a KKT point, hence it is attained for $x_1 = x_2 = \frac{P}{4}$. The rectangle that maximises the area for a given perimeter is a square.

Exercice 3.2

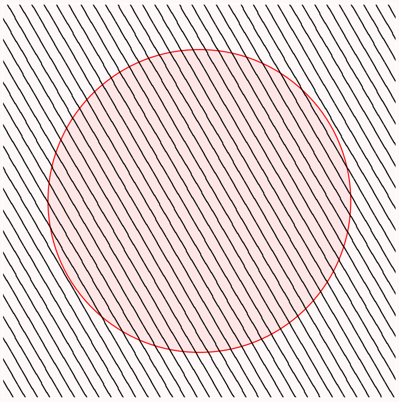
Show that the following problem admits a solution and find it using KKT theorem.

$$\min_{x^2+y^2\leq 1} \; 3x + 2y$$

Solution to Exercice 3.2

We split the solution into several different simple steps.

- **Step −1 : Standard form** Let $M = (x, y) \in \mathbb{R}^2$, define the functions $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ as
$$f(M) = 3x + 2y \quad g(M) = x^2 + y^2 - 1.$$
If $X = \{M \in \mathbb{R}^2 \text{ s.t. } g(M) \leq 0\}$, then the problem reads $\min_X f$.
- **Step 0 : Explanatory Figure** The set X is displayed in red and the level-sets of f are displayed in black.



- **Step 1 : Existence of a minimum** The function f is continuous, the set X is bounded and the set X is closed (because g is continuous).
- **Step 2 : Qualification of constraints** The inequality constraint is convex and $g(0) < 0$. The constraints are qualified everywhere by Slater's condition. We proved that the constraints are qualified at each point.
- **Step 3 : Solving KKT** The Lagrangian is given by
$$\mathcal{L}(M, \lambda) = 3x + 2y + \lambda(x^2 + y^2 - 1).$$

The KKT equations read

$$\begin{cases} (1): & \begin{pmatrix} 3 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2x \\ 2y \end{pmatrix} = 0 \\ (2): & \lambda \geq 0 \text{ and } g(M) \leq 0 \\ (3): & \text{either } \lambda = 0 \text{ or } g(M) = 0 \end{cases}$$

We discuss according to the cases in (3).

- **If $g(M) \neq 0$:** Then we must have $\lambda = 0$ by (3). Equation (1) yields $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 0$ which is a contradiction.
- **If $g(M) = 0$:** Then (3) is verified and (2) boils down to $\lambda \geq 0$. From (1) we can see that $\lambda \neq 0$ and then $x = \frac{-3}{2\lambda}$ and $y = \frac{-1}{\lambda}$. From $g(M) = 0$, we have
$$\frac{9}{4\lambda^2} + \frac{1}{\lambda^2} = 0$$
And we have $\lambda^2 = \frac{13}{4}$. Recalling that $\lambda \geq 0$, we have $\lambda = \frac{\sqrt{13}}{2}$ and $x = \frac{-3}{2\lambda}$ and $y = \frac{-1}{\lambda}$.

We have only one KKT point.

- **Step 4 : Conclusion** Our global reasonning is as follows
  1. There exists a least one global minimizer, it is a local minimizer.
  2. The constraints are qualified everywhere, each local minimizer is a KKT point.
  3. There exists only one KKT point at point $\lambda = \frac{\sqrt{13}}{2}$ and $x = \frac{-3}{2\lambda}$ and $y = \frac{-1}{\lambda}$.
  4. There exists only one local minimizer at point $\lambda = \frac{\sqrt{13}}{2}$ and $x = \frac{-3}{2\lambda}$ and $y = \frac{-1}{\lambda}$. It is a global minimizer.

3.2 Fundamental theorem of linear inequalities

Before tackling the KKT theorem, we take a little detour and state (and prove) a very important lemma.

3.2.1 Statement of the fundamental theorem and explanation

Theorem 3.2.1 — Fundamental theorem of linear inequalities. Let E be an Hilbert space and $(a_i)_{i=1,\dots,m} \in E$ be a finite family of vectors and $b \in E$. Then exactly one of the two following proposition is true

- $\exists c \in \mathbb{R}^m$ such that $b = \sum_{i=1}^m c_i a_i$ and $c_i \geq 0$ for all i.
- $\exists b^\star \in E$ such that $(b, b^\star) < 0$ and $(b^\star, a_i) \geq 0$ for all i.

In other words if we define the **the cone generated by the $(a_i)_i$** as :

$$\mathcal{C} = \left\{ x = \sum_{i=1}^m c_i a_i \text{ such that } c_i \geq 0 \quad \forall i \right\},$$

then either $b \in \mathcal{C}$ or there exist an hyperplane $\mathcal{H}$ that separates b and $\mathcal{C}$. This hyperplane is defined as

$$\mathcal{H} = \{x, \text{ such that } (b^\star, x) = 0\}$$

3.2.2 Proof of Theorem 3.2.1

Theorem **??** is a simplified version of a theorem called **Farka's lemma**. The first easy remark is to prove that the two alternatives in Theorem **??** are mutually exclusive, indeed if they were both true, then we would have

$$0 \leq \sum_{i=1}^{m} c_i(a_i, b^\star) = (\sum_{i=1}^{m} c_i a_i, b^\star) = (b, b^\star) < 0.$$

Which is an obvious contradiction. We then have to show that when the first alternative is false, the second has to be true. The proof relies on the theorem of separation of convex set, it is a very important theorem of convexity, originally due to Minkowski (in finite dimension). It is extended to Banach spaces vector spaces by Hahn and Banach and is known as the Hahn-Banach theorem (geometrical form). We are not in position to prove it in the setting of a Banach space, we prove it for an Hilbert space.

Proposition 3.2.2 — Hahn-Banach in an Hilbert. Let E be an Hilbert space and $\mathcal{C} \subset E$ be a closed, convex, non-empty set and $b \notin \mathcal{C}$, then there exists $b^\star$ and $\alpha \in \mathbb{R}$ such that

$$\langle b^\star, b\rangle < \alpha < \langle b^\star, c\rangle \quad \forall c \in \mathcal{C}.$$

Hence the hyperplane $\mathcal{H} = \{x, \text{ such that } \langle b^\star, x\rangle = \alpha\}$ striclty separates $\mathcal{C}$ and $\{b\}$.

Proof

Denote π the orthogonal projection of b on $\mathcal{C}$, that is π is the unique minimizer of

$$\min_{p\in\mathcal{C}} \frac{1}{2}\|p-b\|^2.$$

We have existence and uniqueness of π by Proposition 2.3.3 and we have, for all $c \in \mathcal{C}$.

$$\langle \pi - b, c - \pi\rangle \geq 0.$$

Denote $b^\star = \pi - b$. Because $b \notin \mathcal{C}$, we have $b^\star \neq 0$ and

$$0 \quad \leq \quad \langle b^\star, c-\pi\rangle = \langle b^\star, c-b+b-\pi\rangle = \langle b^\star, c-b-b^\star\rangle$$

And then

$$\langle b^\star, b\rangle + \|b^\star\|^2 \leq \langle b^\star, c\rangle.$$

It is then sufficient to take $\alpha = \langle b^\star, b\rangle + \frac{1}{2}\|b^\star\|^2$ and to recall that $b^\star \neq 0$ to conclude.

We can now turn to the proof of Theorem 3.2.1.

Proof of Theorem 3.2.1

Denote

$$\mathcal{C} = \left\{ x = \sum_{i=1}^{m} c_i a_i \text{ such that } c_i \geq 0 \quad \forall i \right\},$$

It is easy to prove that $\mathcal{C}$ is a convex set. We first prove that it is closed and we do it by an induction on m. The case $m = 1$ is trivial. Now take $(w_n)_n$ a converging sequence of elements of C towards some $w \in E$. We want to prove that $w \in \mathcal{C}$

- If the $(a_i)_i$ are linearly independent, then the decomposition $w_n = \sum_i c_i^n a_i$ is unique and convergence of $(w_n)_n$ is equivalent ot the convergence of $(c^n)_n$ and then $(c^n)_n$ converges to $c \succeq 0$ and $w = \sum_i c_i^n a_i \in C$.
- Suppose now that the $(a_i)_i$ are not linearly independent, there exist a linear combination of the form $\sum_i \mu_i a_i = 0$. Take any $\omega = \sum_i c_i a_i \in C$. Because $c \succeq 0$, there exists a small t such that $c + t\mu \succeq 0$ and for at least one index i_0 we have $c_{i_0} + t\mu_{i_0} = 0$. It follows that

$$\omega = \sum_i c_i a_i + t \sum_i \mu_i a_i = \sum_{i \neq i_0} (c_i + t\mu_i) a_i.$$

Hence we have proven that

$$\mathcal{C} \subset \bigcup_{i_0=1}^{p} \left\{ v \in \mathbb{R}^n, \exists \gamma \succeq 0 \text{ such that } v = \sum_{i \neq i_0} \gamma_i a_i \right\}.$$

Each of the set on the right-hand side is closed (by the recurrence hypothesis) and hence the same is true for $\mathcal{C}$.

Now either $b \in \mathcal{C}$ or $b \notin \mathcal{C}$, in the later case, by Hahn-Banach theorem, there exists $b^\star$ and α such that

$$\langle b^\star, b \rangle < \alpha < \langle b^\star, c \rangle \quad \forall c \in \mathcal{C}. \tag{3.1}$$

Because $0 \in \mathcal{C}$, when we put $c = 0$ in (3.1), we obtain $\alpha < 0$ and then $\langle b^\star, b \rangle < 0$. For any i and $t \geq 0$, consider $c = ta_i$, then $c \in \mathcal{C}$. Set $c = ta_i$ in (3.1) and obtain $\alpha < t(b^\star, a_i)$. Sending t to $+\infty$ shows that $(b^\star, a_i) < 0$ is impossible, so that $(b^\star, a_i) \geq 0$.

3.3 The tangent cone

3.3.1 Definition

The tangent cone is a very important tool in the analysis of the behavior of a function nears its local optima.

Definition 3.3.1 — Tangent cone. A direction d is tangent to X in $x \in X$ iff there exists a sequence $(d_n)_{n \in \mathbb{N}}$ that converges to d and a sequence of real positive numbers $(\varepsilon_n)_{n \in \mathbb{N}}$ that converges to 0 such that :

$$\text{For all } n, \quad x + \varepsilon_n d_n \in X.$$

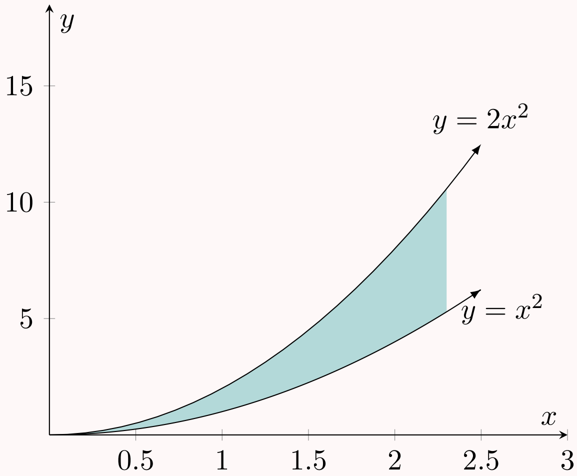
We denote $T_x(X)$ the set of tangent directions of X at point x. This set $T_x(X)$ is called the tangent cone of X at point x. Equivalently, denoting $x_n = x + \varepsilon_n d_n$, we have

$$T_x(X) = \left\{ d \text{ s. t. } \exists (x_n, \varepsilon_n)_n \in (X \times \mathbb{R}^{+*})^{\mathbb{N}} \text{ with } (x_n, \varepsilon_n, \frac{x_n - x}{\varepsilon_n}) \to (x, 0, d) \right\}$$

Exercice 3.3

Let $X = \{x \in \mathbb{R}^2 | x_1^2 \leq x_2 \leq 2x_1^2 \text{ and } x_1 \geq 0\}$. Draw X in $\mathbb{R}^2$ and compute $T_{(0,0)}(X)$.

Solution to Exercice 3.3



- We show that the cone $T_{(0,0)}(X)$ is reduced to $\lambda(1,0)$, with $\lambda \in \mathbb{R}$. Suppose that $d = (a,b)$ with $b \neq 0$ belongs to the cone, then there exist $x_n = \varepsilon_n d_n \in X$ such that $(x_n, \varepsilon_n, d_n)$ converges to $(0,0,(a,b))$. We write $d_n = (a_n, b_n)$ and we must have
$$\varepsilon_n^2 a_n^2 \leq \varepsilon_n b_n \leq 2\varepsilon_n^2 a_n^2.$$
Divide by ε_n and let n goes to $+\infty$ to obtain $b = 0$.
- We show that the cone $T_{(0,0)}(X)$ is reduced to $\lambda(1,0)$, with $\lambda \in \mathbb{R}^+$. Indeed if it where to be the case, there would exists $\varepsilon_n > 0$ and $d_n = (a_n, b_n)$ such that $a_n \varepsilon_n \geq 0$ for every n and a_n converges to $\lambda < 0$, which is impossible.
- We now show that $(1,0)$ belongs to $T_{(0,0)}(X)$. Let $\varepsilon_n = \frac{1}{n}$ and $d_n = (1, \frac{1}{n})$ we have $x_n = (\frac{1}{n}, \frac{1}{n^2})$ belongs to X for every n and converges to $(0,0)$.

3.3.2 Inequality and equality constraints

We are now interested in characterizing the tangent cone of a set given by equality and inequality constraints in standard form by :

$$X = \{x \in \mathbb{R}^n \text{ such that } g_i(x) = 0 \forall i \in \mathcal{E}, g_i(x) \leq 0 \forall i \in \mathcal{I}\},$$

We recall the definition of active constraints :

Definition 3.3.2 — Active constraints. Denote $\mathcal{A}_x$ the set of indices $i \in \mathcal{I}$ such that $g_i(x) = 0$. This set is called the **set of active constraints at** x.

$$\mathcal{A}_x = \{i \in \mathcal{I} \text{ such that } g_i(x) = 0\}$$

Using Taylor's expansions, it is easy to give a necessary conditions for a direction d to belong to the tangent cone.

Proposition 3.3.1 Suppose that g_i is differentiable for all i, then for all $d \in T_x(X)$ we have :

- For all $i \in \mathcal{E}$, then $\langle \nabla g_i(x), d \rangle = 0$
- For all $i \in \mathcal{A}_x$, then $\langle \nabla g_i(x), d \rangle \leq 0$

Proof

Let $d \in T_x(X)$, then there exists $(d_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ with respective limits d and 0 s.t.

$$x_n = x + \varepsilon_n d_n \in X.$$

By a Taylor expansion:

$$g_i(x_n) = g_i(x) + \langle \nabla g_i(x), \varepsilon_n d_n \rangle + \mathcal{O}(\varepsilon_n d_n)$$

- If $i \in \mathcal{E}$, then $g_i(x) = g_i(x_n) = 0$ and we have

$$\langle \nabla g_i(x), \varepsilon_n d_n \rangle + \mathcal{O}(\varepsilon_n d_n) = 0$$

- If $i \in \mathcal{A}_x$, then $g_i(x) = 0$ and $g_i(x_n) \leq 0$ and we have

$$\langle \nabla g_i(x), \varepsilon_n d_n \rangle + \mathcal{O}(\varepsilon_n d_n) \leq 0$$

Finish by dividing by ε_n and letting n goes to $+\infty$ (then d_n goes to d).

Thanks to the above proposition, we know that the tangent cone obeys the following inclusion :

$$T_x(X) \subset \{d \text{ such that } \langle \nabla g_i(x), d \rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x), d \rangle \leq 0 \ \forall i \in \mathcal{A}_x\}$$

Sadly, in order to prove KKT, we need the reversed inclusion. So we put this reversed inclusion as an hypothesis, and we call this hypothesis the qualification of constraints.

Definition 3.3.3 We say that **the constraints are qualified at point x** if

$$T_x(X) = \{d \text{ such that } \langle \nabla g_i(x), d \rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x), d \rangle \leq 0 \ \forall i \in \mathcal{A}_x\}$$

3.4 Proof of KKT theorem

3.4.1 Main proof

Proposition 3.4.1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^\star$ a local minimum of f on X, if the constraints are qualified at point $x^\star$, then there exist λ_i such that $\lambda_i \geq 0$ for all $i \in \mathcal{I}$ and

$$\nabla f(x^\star) + \sum_i \lambda_i \nabla g(x^\star) = 0 \text{ and } \lambda_i g_i(x^\star) = 0 \quad \forall i$$

Proof

Take the family $(\nabla g_i(x^\star))_{i\in\mathcal{E}} \cup (-\nabla g_i(x^\star))_{i\in\mathcal{E}} \cup (\nabla g_i(x^\star))_{i\in\mathcal{A}_x}$ and denote this family

$(a_j)_j$. The equations

$$\nabla f(x^\star) + \sum_i \lambda_i \nabla g(x^\star) = 0 \text{ and } \lambda_i g_i(x^\star) = 0 \quad \forall i \text{ and } \lambda_i \geq 0 \; \forall i \in \mathcal{I}$$

is equivalent to saying that there exists $c_j \geq 0$ such that $-\nabla f(x) = \sum_j c_j a_j$. Suppose now that there is no such $(c_j)_j$ by the fundamental Theorem 3.2.1, there exists $d^\star$ such that $(-\nabla f(x^\star), d^\star) > 0$ and

$$\langle \nabla g_i(x^\star), d^\star \rangle = 0 \; \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x^\star), d^\star \rangle \leq 0 \; \forall i \in \mathcal{A}_x.$$

By the definition of the qualification of constraints, it means that $d^\star \in T_{x^\star}(X)$. Introduce $(d_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$ with :

$$x_n = x^\star + \varepsilon_n d_n \in X, \quad \lim d_n = d^\star, \quad \lim \varepsilon_n = 0$$

Since $x^\star$ is a local minimum of f on X, there exists $r > 0$ such that:

$$\forall x \in B(x^\star, r) \cap X, \quad f(x) \geq f(x^\star),$$

Hence there exist N s.t. $f(x_n) \geq f(x^\star)$, for $n \geq N$. By Taylor :

$$f(x_n) = f(x^\star + \varepsilon_n d_n) = f(x^\star) + \varepsilon_n \langle \nabla f(x^\star), d_n \rangle + \mathcal{O}(\varepsilon_n d_n)$$

Hence: $\varepsilon_n \langle \nabla f(x^\star), d_n \rangle + \mathcal{O}(\varepsilon_n d_n) \geq 0$, and dividing by ε_n and letting n goes to $+\infty$ yields $\langle \nabla f(x^\star), d^\star \rangle \geq 0$ which is a contradiction.

3.4.2 Proving qualification of constraints : LICQ

The qualification of constraint is an abstract hypothesis given in Definition 3.3.3. In this section, we prove that the LICQ condition ensures constraints qualification. The proof is based on the Implicit Function Theorem that we recall below

Proposition 3.4.2 — Implicit Function Theorem. Let $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be a C^1 function. Let $x^\star$ be a solution of $h(x^\star, 0) = 0$. Denote $D_x h(x^\star, 0) \in \mathcal{M}_{n,n}(\mathbb{R})$ the Jacobian of h with respect to x only at point $(x^\star, 0)$. Suppose that $D_x h(x^\star, 0)$ is invertible, there exists $\mathcal{U}$ a neighbourhood of $x^\star$ and $\mathcal{V}$ a neighbourhood of $0 \in \mathbb{R}$ and a mapping $\gamma : \mathcal{V} \to \mathcal{U}$ such that

$$\Big(h(x,t) = 0 \text{ and } (x,t) \in (\mathcal{U}, \mathcal{V})\Big) \Longleftrightarrow x = \gamma(t).$$

Moreover $\dot\gamma(t) = \frac{\partial}{\partial t} x(t)$ verifies

$$D_x h(\gamma(t), t)\dot\gamma(t) + D_t h(\gamma(t), t) = 0$$

Proposition 3.4.3 — LICQ. Suppose that X is in standard form and suppose that the family

$$(\nabla g_j(x^\star))_{j \in \mathcal{E} \cup \mathcal{A}_{x^\star}}$$

is linearly independent. Then the constraints are qualified at $x^\star$, that is, if d verifies

$$\langle \nabla g_i(x^\star), d\rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x^\star), d\rangle \le 0 \ \forall i \in \mathcal{A}_{x^\star} \tag{3.2}$$

Then $d \in T_{x^\star}(X)$. It is even possible to find a sequence $(\varepsilon_n, d_n) \in \mathbb{R}^+ \times \mathbb{R}^n$ that converges to $(0, d)$ with $x_n = x^\star + \varepsilon_n d_n \in X$ and

$$g_i(x_n) = \varepsilon_n \langle \nabla g_j(x^\star), d\rangle \text{ for every } i \in \mathcal{E} \cup \mathcal{A}_{x^\star}. \tag{3.3}$$

Proof

We consider $\mathcal{J} = \mathcal{E} \cup \mathcal{A}_{x^\star}$, and $J = |\mathcal{J}|$ the cardinality of $\mathcal{J}$. Since $(\nabla g_j(x^\star))_{j\in\mathcal{J}}$, is linearly independent, we can find vectors $(z_j)_{j=J+1,\dots n}$ such that $(\nabla g_j(x^\star))_{j\in\mathcal{J}} \cup (z_j)_j$ is a basis of $\mathbb{R}^n$. We denote $G(x) \in \mathbb{R}^J$ and $Z \in \mathcal{M}_{n-J,J}(\mathbb{R})$ by :

$$G(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_J(x) \end{pmatrix} \text{ and } Z = \begin{pmatrix} z_{J+1}^T \\ \vdots \\ z_n^T \end{pmatrix}.$$

We consider $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be defined as

$$h(x,t) = \begin{pmatrix} G(x) - t(DG)(x^\star)d \\ Z(x - x^\star - td) \end{pmatrix}$$

Then $Dh = \begin{pmatrix} D_x h & D_t h \end{pmatrix}$, with

$$D_x h(x^\star, 0) = \begin{pmatrix} (DG)(x^\star) \\ Z \end{pmatrix} \text{ and } D_t h(x^\star, 0) = \begin{pmatrix} -(DG)(x^\star)d \\ -Zd \end{pmatrix}$$

Since $h(x^\star, 0) = 0$ and $D_x h(x^\star, 0)$ is invertible, the Implicit Function Theorem 3.4.2 states that there exists a C^1-curve $t \mapsto x(t)$ and $\mathcal{V}$, a neighbourhood of $x^\star$ such that for every small t and for each $x \in \mathcal{V}$:

$$h(x,t) = 0 \Leftrightarrow x = x(t).$$

We will take any sequence ε_n that tends to 0 and $x_n = x(\varepsilon_n)$ In particular, we have $x(0) = x^\star$ and $G(x(t)) = tDG(x^\star)d$, that is

$$g_j(x(t)) = t\langle \nabla g_j(x^\star), d\rangle \text{ for all } j \in \mathcal{J}. \tag{3.4}$$

In particular, (3.3) is (3.4) in disguise. Equation (3.4) and (3.2) yield $g_i(x(t)) = 0$ for all $i \in \mathcal{E}$ and $g_i(x(t)) \le 0$ for all $i \in \mathcal{A}_{x^\star}$ and $t \ge 0$. For all $i \in \mathcal{I} \setminus A_{x^\star}$, we have $g_i(x^\star) < 0$, hence $g_i(x(t)) < 0$ for small $t \ge 0$. This allows to conclude that $x(t) \in X$ for every small t. We need to show that $\frac{x(t) - x^\star}{t}$ converges to d as t goes to 0. Using the implicit function theorem, we have

$$D_x h(x^\star, 0)\frac{d}{dt}x(0) + D_t h(x^\star, 0) = 0 \Longrightarrow \frac{d}{dt}x(0) = -(D_x h(x^\star, 0))^{-1} D_t h(x^\star, 0) = d.$$

3.4.3 Proving qualification of constraints : Slater

Before that Slater's condition yields constraint qualification, we have to prove that the tangent cone is a closed cone.

Proposition 3.4.4 For any $x \in X$, then $T_x(X)$ is a closed cone.

Proof

If d in $T_x(X)$. For any $\lambda > 0$, replace ε_n by $\frac{\varepsilon_n}{\lambda}$ to prove that λd is in $T_x(X)$, hence $T_x(X)$ is a cone. If d_m is a sequence in $T_x(X)$ that converges to some d we perform a diagonal sequence argument to show that d in $T_v(K)$.

- For every fixed m, we have a sequence $M_{mn} = (x_{mn}, \varepsilon_{mn}, \frac{x_{mn}-x}{\varepsilon_{mn}})$ that converges as n goes to $+\infty$ towards $M_m = (x, 0, d_m)$. Moreover M_m converges as m goes to $+\infty$ towards $M = (x, 0, d)$. The goal is to find a subsequence of M_{mn} that converges to M.
- Take $p \in \mathbb{N}$, let m be a large integer so that $\|M_m - M\| \leq \frac{1}{2p}$. Now let n be some large integer so that $\|M_m - M_{mn}\| \leq \frac{1}{2p}$, we have $\|M_{mn} - M\| \leq \frac{1}{p}$. Repeat the process for every p, we have a sequence of M_{mn} that converges to M.

Proposition 3.4.5 — Slater's conditions. Suppose that for each $i \in \mathcal{I}$, g_i is convex and forall $i \in \mathcal{E}$, g_i is affine. If there exists a point $x_0 \in X$ such that for each i

$$g_i(x_0) < 0 \text{ or } g_i \text{ is affine,}$$

then the constraints are qualified everywhere in X.

Proof

Let $x^\star \in X$ and d be such that

$$\langle \nabla g_i(x^\star), d\rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x^\star), d\rangle \leq 0 \ \forall i \in \mathcal{A}_{x^\star} \tag{3.5}$$

The goal is to prove that $d \in T_{x^\star}(X)$. Denote $w = x_0 - x^\star$. For every $t \in [0,1]$, $x^\star + tw = tx_0 + (1-t)x^\star \in X$ and consider $x_n = x^\star + \varepsilon_n(d + tw)$ for some sequence ε_n that goes to 0 as n goes to $+\infty$.

- If $i \in \mathcal{I} \setminus \mathcal{A}_{x^\star}$, then $g_i(x^\star) < 0$ and hence $g_i(x_n) < 0$ for large n.
- If $i \in A_{x^\star}$ and not affine, then $g_i(x_0) < 0$ and

  $$\langle \nabla g_i(x^\star), w\rangle = \langle \nabla g_i(x^\star), x_0 - x^\star\rangle \leq g_i(x_0) - g_i(x^\star) = g_i(x_0) < 0$$

  Hence

  $$g_i(x_n) = g_i(x^\star) + \varepsilon_n\langle \nabla g_i(x^\star), d + tw\rangle + o(\varepsilon_n) \leq \varepsilon_n t\langle \nabla g_i(x^\star), w\rangle + o(\varepsilon_n)\,.$$

  We then have $g_i(x_n) \leq 0$ for large n.
- If g_i is affine and $g_i(x^\star) = 0$, then

  $$\langle \nabla g_i(x^\star), w\rangle = \langle \nabla g_i(x^\star), x_0 - x^\star\rangle = g_i(x_0) - g_i(x^\star) = g_i(x_0)$$

  Hence

  $$g_i(x_n) = g_i(x^\star) + \varepsilon_n\langle \nabla g_i(x^\star), d + tw\rangle = \varepsilon_n t\langle \nabla g_i(x^\star), d\rangle + t\varepsilon_n g_i(x_0).$$

If $i \in \mathcal{E}$, then $\langle \nabla g_i(x^\star), d\rangle = 0$ and $g(x_0) = 0$ and then $g_i(x_n) = 0$. If $i \in \mathcal{I}$, then $\langle \nabla g_i(x^\star), d\rangle \leq 0$ and $g_i(x_0) \leq 0$ and then $g_i(x_n) \leq 0$.

We have shown that $x_n \in X$ for large n. Hence $d + tw \in T_{x^\star}(X)$ for every small t. Using the closedness of $T_{x^\star}(X)$ then $d \in T_{x^\star}(X)$.

3.5 More on KKT Theorem

3.5.1 Transforming inequalities into equalities

We recall that if we can alternatively define X with inequalities only as :

$$X = \{x \text{ such that } g_i(x) \leq 0, \forall i \in \mathcal{I} \cup \mathcal{E} \text{ and } -g_i(x) \leq 0 \forall i \in \mathcal{E}\}$$

In this case, we study how the different definitions/theorem change :

- The LICQ condition changes into the following : the family

$$(\nabla g_i(x))_{i\in\mathcal{I}} \cup (\nabla g_i(x))_{i\in\mathcal{E}} \cup (-\nabla g_i(x))_{i\in\mathcal{E}}$$

  must be linearly independent. It is impossible to fulfill this condition if $\mathcal{E}$ is not empty.
- The Slater's condition is : For each $i \in \mathcal{E}$, then g_i and $-g_i$ must be convex. This is equivalent to saying that g_i is affine. There is no other condition. Hence Slater's condition is exactly the same.
- The Lagrangian : We have a new Lagrangian $\tilde{\mathcal{L}}(x, \tilde{\lambda})$ which is defined as

$$\tilde{\mathcal{L}}(x, \tilde{\lambda}) = f(x) + \sum_{i\in I} \tilde{\lambda}_i g_i(x) + \sum_{i\in E} \tilde{\lambda}_i^+ g_i(x) - \sum_{i\in E} \tilde{\lambda}_i^- g_i(x),$$

  where $\tilde{\lambda} = ((\tilde{\lambda}_i)_{i\in\mathcal{I}}, (\tilde{\lambda}_i^+)_{i\in\mathcal{E}}, (\tilde{\lambda}_i^-)_{i\in\mathcal{E}})$. We see that upon setting

$$\lambda = ((\tilde{\lambda}_i)_{i\in\mathcal{I}}, (\tilde{\lambda}_i^+ - \tilde{\lambda}_i^-)_{i\in\mathcal{E}}),$$

  we have

$$\mathcal{L}(x, \lambda) = \tilde{\mathcal{L}}(x, \tilde{\lambda}).$$

- KKT conditions are the same.

CHAPTER 4

Second order conditions

4.1 Second order optimality conditions with constraints

As in Chapter 3, we start by stating the conditions in the case where the set X is open.

Proposition 4.1.1 Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a C^2 function, and $X \subset \mathbb{R}^n$ is an open set, then

1. If $x^\star \in X$ is a local minimum of f on X then $\nabla f(x^\star) = 0$ and $H[f](x^\star) \succeq 0$.
2. If $x^\star \in X$ is a point such that $\nabla f(x^\star) = 0$ and $H[f](x^\star) \succ 0$ then $x^\star$ is a local minimum of f over X.

Proof

We recall the formula of the second-order Taylor expansion, valid here since f is a C^2 function. For each point $x^\star \in X$ and direction $d \in \mathbb{R}^n$, for each $\varepsilon > 0$ small enough, then $x^\star + \varepsilon d \in X$ (because X is an open set) and we have

$$f(x^\star + \varepsilon d) = f(x^\star) + \varepsilon\langle \nabla f(x^\star), d\rangle + \frac{\varepsilon^2}{2}\langle H[f](x^\star)d, d\rangle + o\big(\varepsilon^2\big)$$

1. First suppose that $x^\star \in X$ is a local minimum of f on X, then $\nabla f(x^\star) = 0$ by Proposition 3.1.1. Then for ε small enough, we have $f(x^\star + \varepsilon d) - \geq f(x^\star)$, this implies that for any small ε, we have :

$$\frac{\varepsilon^2}{2}\langle H[f](x^\star)d, d\rangle + o\big(\varepsilon^2\big) \geq 0.$$

This implies that it is impossible to find a d such that $\langle H[f](x^\star)d, d\rangle < 0$. In other words, for every d, we must have $\langle H[f](x^\star)d, d\rangle \geq 0$ or $H[f](x^\star) \succeq 0$.

2. Now suppose that $x^\star \in X$ is a point such that $\nabla f(x^\star) = 0$ and $H[f](x^\star) \succ 0$, then there exists a constant $c > 0$ such that $\langle H[f](x^\star)d, d\rangle \geq c\|d\|^2$ for any direction d. And then the Taylor expansions yields

$$f(x^\star + \varepsilon d) \geq f(x^\star) + \frac{\varepsilon^2}{2}c\|d\|^2 + o\big(\varepsilon^2\big)\,.$$

Restricting the directions d to be of unit norm, the constant in the $o()$ does not depend on the direction and we chose ε small enough so that $o\big(\varepsilon^2\big) \geq -\frac{c}{4}\varepsilon^2$.

We then have for each direction d and ε small enough :

$$f(x^\star + \varepsilon d) \geq f(x^\star) + \frac{\varepsilon^2}{4} c\|d\|^2.$$

Which is exactly the defition of $x^\star$ being a local minimum.

In the case where X is defined through equality and inequality constraints, the theorem has the same flavor, the expression of the theorem is a little bit more involved. We first recall the notion of a KKT point

Definition 4.1.1 — KKT point. Let f, and g_i be differentiable functions. We have the following definitions

- The set of constraints X is given by

$$X = \{x \in E \text{ such that } g_i(x) \leq 0 \quad \forall i \in \mathcal{I} \text{ and } g_i(x) = 0 \quad \forall i \in \mathcal{E}\}.$$

- The Lagrangian $\mathcal{L}$ is defined by $\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x)$, the set Λ is defined by

$$\Lambda = \{\lambda \in \mathbb{R}^p \text{ such that } \lambda_i \geq 0, \ \forall i \in \mathcal{I}\}$$

- A point $M = (x, \lambda)$ is said to be a KKT point if

$$\nabla_x \mathcal{L}(M) = 0, \quad x \in X, \quad \lambda \in \Lambda \text{ and } g_i(x)\lambda_i = 0 \quad \forall i$$

We now state our main theorem on second order conditions for sets X that are in standard form.

Theorem 4.1.2 — 2nd order conditions.

- At each point $M = (x, \lambda)$, the **linearising cone of constraints** is defined by

$$V_M = \left\{ d \in \mathbb{R}^n \text{s.t.} \begin{cases} \langle \nabla g_i(x), d\rangle = 0 & \forall i \in \mathcal{E} \\ \langle \nabla g_i(x), d\rangle \leq 0 & \forall i \in \mathcal{A}_x \\ \lambda_i \langle \nabla g_i(x), d\rangle = 0 & \forall i \end{cases} \right\}$$

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^p$, be C^2 functions and let

$$X = \{x \in \mathbb{R}^n \text{ such that } g_i(x) = 0 \ \forall i \in \mathcal{E} \text{ and } g_i(x) \leq 0 \ \forall i \in \mathcal{I}\}.$$

1. **Necessary :** If the LICQ conditions hold at $x^\star \in X$ and $x^\star \in X$ is a local minimizer of f on X then there exists $\lambda^\star$ such that $M^\star = (x^\star, \lambda^\star)$ is a KKT point and
$$\forall d \in V_{M^\star} \quad \langle H_x[\mathcal{L}](M^\star)d, d\rangle \geq 0.$$
2. **Sufficient :** If there exists $\lambda^\star$ such that $M^\star = (x^\star, \lambda^\star)$ is a KKT point and if
$$\forall d \in V_{M^\star} \text{ with } d \neq 0 \text{ then} \quad \langle H_x[\mathcal{L}](M^\star)d, d\rangle > 0,$$
then $x^\star$ is a local minimizer of f over X.

We can check that if an equality constraint is modified into two inequality constraints, then the linearising cone of constraints is unchanged. We only discussed the LICQ condition in the necessary condtion, indeed if one verifies the Slater condtion, then the Lagrangien is a convex C^2 function and we always have $H_x[\mathcal{L}] \succeq 0$. Moreover we will prove later than for a convex problem a KKT point is always a minimum of the original problem. This proves the sufficient second order condition for Slater's case.

4.1.1 Proof of the necessary second order condition

Let $x^\star$ be a local minimizer of f over X and suppose that LICQ conditions holds at $x^\star$, there exists $M^\star = (x^\star, \lambda^\star)$ such that $M^\star$ is a KKT point (by Theorem 3.1.4). Suppose that there exists $d \in V(M^\star)$ with $\langle H_x[\mathcal{L}](M^\star)d, d\rangle < 0$. We have in particular

$$\langle \nabla g_i(x^\star), d\rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x^\star), d\rangle \leq 0 \ \forall i \in \mathcal{A}_{x^\star} \tag{4.1}$$

By Proposition 3.4.3, we have $d \in T_{x^\star}(X)$. It is even possible to find a sequence $(\varepsilon_n, d_n) \in \mathbb{R}^+ \times \mathbb{R}^n$ that converges to $(0, d)$ with $x_n = x^\star + \varepsilon_n d_n \in X$ and

$$g_i(x_n) = \varepsilon_n \langle \nabla g_i(x^\star), d\rangle \text{ for every } j \in \mathcal{E} \cup \mathcal{A}_{x^\star}. \tag{4.2}$$

Now recall that $\lambda_i^\star = 0$ if $i \notin \mathcal{E} \cup \mathcal{A}_{x^\star}$, we have

$$\mathcal{L}(x_n, \lambda^\star) = f(x_n) + \sum_{i \in \mathcal{E} \cup \mathcal{A}_{x^\star}} \lambda_i^\star g_i(x_n) = f(x_n) + \sum_{i \in \mathcal{E} \cup \mathcal{A}_{x^\star}} \varepsilon_n \lambda_i^\star \langle \nabla g_i(x^\star), d\rangle = f(x_n)$$

We then perform a second order expansion of the mapping $x \mapsto \mathcal{L}(x, \lambda^\star)$ around $x^\star$, we have

$$\begin{aligned} f(x_n) &= \mathcal{L}(x^\star + \varepsilon_n d_n, \lambda^\star) = f(x^\star) + \varepsilon_n \langle \underbrace{\nabla_x \mathcal{L}(M^\star)}_{=0 \text{ (KKT)}}, d_n\rangle + \frac{\varepsilon_n^2}{2} \langle H_x[\mathcal{L}](M^\star)d_n, d_n\rangle + o(\varepsilon_n^2) \\ &= f(x^\star) + \frac{\varepsilon_n^2}{2}\Big(\langle H_x[\mathcal{L}](M^\star)d_n, d_n\rangle + o(1)\Big) \end{aligned}$$

Because $x^\star$ is a local minimum of f over X, we conclude that for large n, we must have

$$\langle H_x[\mathcal{L}](M^\star)d_n, d_n\rangle + o(1) \geq 0.$$

Letting n go to infinity, we are in contradiction with $\langle H_x[\mathcal{L}](M^\star)d, d\rangle < 0$.

4.1.2 Proof of the sufficient second order condition

We suppose $M^\star = (x^\star, \lambda^\star)$ is a KKT point of f and that

$$\forall d \in V_{M^\star} \text{ with } d \neq 0 \text{ then } \quad \langle H_x[\mathcal{L}](M^\star)d, d\rangle > 0,$$

In order to prove that $x^\star$ is a local minimum of f, we suppose it is not the case and we show a contradiction. If $x^\star$ is not a local minimum, there exists a sequence of $x_n \in X$ that converges to $x^\star$ such that $f(x_n) \leq f(x^\star)$ and $x_n \neq x^\star$. Denote $\varepsilon_n = \|x_n - x^\star\|$ and $d_n = \frac{x_n - x^\star}{\|x_n - x^\star\|}$. The sequence d_n is bounded and converges up to a subsequence. Then up to a subsequence (still denoted $\bullet_n$), we have $x_n = x + \varepsilon_n d_n$ and d_n converges to some d and $\varepsilon_n > 0$ converges to 0. Hence $d \in T_{x^\star}(X)$ and in particular

$$\langle \nabla g_i(x^\star), d\rangle = 0 \ \forall i \in \mathcal{E} \text{ and } \langle \nabla g_i(x^\star), d\rangle \leq 0 \ \forall i \in \mathcal{A}_{x^\star} \tag{4.3}$$

But we have

$$f(x^\star) \geq f(x_n) = f(x^\star) + \varepsilon_n \langle \nabla f(x^\star), d_n\rangle + o(\varepsilon_n)\,.$$

We then have $\langle \nabla f(x^\star), d_n\rangle + o(1) \leq 0$ and letting n go to infinity we must have $\langle \nabla f(x^\star), d\rangle \leq 0$, using the fact that $M^\star$ is a KKT point, we have $\nabla f(x^\star) = -\sum_i \lambda_i \nabla g_i(x^\star)$ and must have

$$\langle \sum_i \lambda_i \nabla g_i(x^\star), d\rangle \geq 0. \tag{4.4}$$

Combining (4.3) and (4.4), we obtain that $d \in V(M^\star)$ and $d \neq 0$ hence we have $\langle H_x[\mathcal{L}](M^\star)d, d\rangle > 0$.

On the other hand, we have

$$\begin{aligned} f(x^\star) &\geq f(x_n) \geq f(x_n) + \sum_i \lambda_i^\star g_i(x_n) = \mathcal{L}(x_n, \lambda^\star) \\ &= \underbrace{\mathcal{L}(M^\star)}_{=f(x^\star)} + \varepsilon_n \underbrace{\langle \nabla_x \mathcal{L}(M^\star)}_{=0}, d_n\rangle + \frac{\varepsilon_n^2}{2}\langle H_x[\mathcal{L}](M^\star)d_n, d_n\rangle + o(\varepsilon_n^2) \end{aligned}$$

This implies that $\langle H_x[\mathcal{L}](M^\star)d_n, d_n\rangle + o(1) \leq 0$ which is in contradiction with $\langle H_x[\mathcal{L}](M^\star)d, d\rangle > 0$. Hence there is no such sequence $(x_n)_n$ and $x^\star$ is a local (strict) minimum

4.2 Interlude : a little exercise

In this section we apply the different theorems in an exercise.

Exercice 4.1

Let $a \in \mathbb{R}^n$, with $a \neq 0$ and $n > 1$. Let Y be the set

$$Y = \{x \in \mathbb{R}^n \text{ s.t. } \|x-a\| = \|a\| \text{ and } \|x - 2a\| = \|a\|\}.$$

Sketch Y and show that

$$Y = \{\frac{3}{2}a + r, \text{ s.t. } (a|r) = 0 \text{ and } \|r\|^2 = \frac{3}{4}\|a\|^2\}.$$

Consider now the problem

$$\min_{x\in X} \|x\| \text{ if } X = \{x \in \mathbb{R}^n \text{ s.t. } \|x-a\| \geq \|a\| \text{ and } \|x-2a\| \leq \|a\|\}$$

1. Show that the problem admits a solution and that the constraints are qualified everywhere.
2. Find every KKT point.
3. Using second order information decide which points are local minimizers and global minimizers.

Solution to Exercice 4.1

Let $x \in Y$, and decompose x into $x = \alpha a + r$ with r orthogonal to a. We have

$$\begin{cases} \|x-a\|^2 = \|a\|^2 \\ \|x-2a\|^2 = \|a\|^2 \end{cases} \implies \begin{cases} |\alpha-1|^2\|a\|^2 + \|r\|^2 = \|a\|^2 \\ |\alpha-2|^2\|a\|^2 + \|r\|^2 = \|a\|^2 \end{cases}$$

From $|\alpha - 1| = |\alpha - 2|$, we deduce $\alpha = \frac{3}{2}$ and then $\|r\|^2 = \frac{3}{4}\|a\|^2$. We follow the standard procedure

- **Step -1 : Standard form** We choose $g_1(x) = \frac{1}{2}\|a\|^2 - \frac{1}{2}\|x-a\|^2$ and $g_2(x) = \frac{1}{2}\|x-2a\|^2 - \frac{1}{2}\|a\|^2$ and $X = \{g(x) \preceq 0\}$. We set $f(x) = \frac{1}{2}\|x\|^2$ and we minimize f over X.

- **Step 0 : Sketch**
- **Step 1 : Existence** The set X is bounded and closed and f is continuous.
- **Step 2 : Qualification of constraints** We discuss according to the number of active constraints
  - no active constraints : nothing to do
  - 1 active constraint : Suppose it is g_1. Then $\nabla g_1(x) = -x + a$ and $\nabla g_1(x) = 0$ implies $x = a$ which implies $g_1(x) \neq 0$. The case for g_2 is handled the same way.
  - 2 active constraints : Then $x \in Y$ and x is written as $x = \frac{3}{2}a + r$ with $(a|r) = 0$ and $\|r\|^2 = \frac{3}{4}\|a\|^2$. The family $(\nabla g_1(x), \nabla g_2(x))$ is the family $(-x0a, x - 2a)$ which is the family $(-\frac{a}{2} - r, -\frac{a}{2} + r)$. Since $(a|r) = 0$ and $a \neq 0$ and $r \neq 0$ this family is linearly independent.

  The constraints are qualified at each point of X.
- **Step 3 : KKT** We first write the Lagrngian which is given by

$$\mathcal{L}(x,\lambda) = \frac{1}{2}\|x\|^2 + \frac{\lambda_1}{2}\left(\|a\|^2 - \|x-a\|^2\right) + \frac{\lambda_2}{2}\left(\|x-2a\|^2 - \|a\|^2\right)$$

The main KKT equation is $\nabla_x \mathcal{L}(x,\lambda) = 0$ which is

$$x - \lambda_1(x-a) + \lambda_2(x-2a) = 0 \tag{4.5}$$

We begin the discussion

- If $\lambda_1 = \lambda_2 = 0$, then (4.5) yields $x = 0$. But $g_2(0) > 0$.
- If $\lambda_1 = 0$ and $\lambda_2 \neq 0$, then (4.5) yields $(1+\lambda_2)x = 2\lambda_2 a$. The case $\lambda_2 = -1$ is impossible and then $x = 2\frac{\lambda_2}{1+\lambda_2}a$. We use :

$$\|x - 2a\| = \frac{2}{|1+\lambda_2|}\|a\|.$$

  The equation $g_2(x) = 0$ and $\lambda_2 \geq 0$ gives $\lambda_2 = 1$ and then $x = a$. But in this case $g_1(x) > 0$. And there is no KKT point.
- If $\lambda_2 = 0$ and $\lambda_1 \neq 0$, then (4.5) yields $(1-\lambda_1)x = \lambda_1 a$. The case $\lambda_1 = 1$ is impossible and then $x = \frac{-\lambda_1}{1-\lambda_1}a$. We use :

$$\|x - a\| = \frac{1}{|1-\lambda_1|}\|a\|$$

  and $g_1(x) = 0$ yields $\lambda_1 = 2$ (the case $\lambda_1 = 0$ is impossible). We check that for the choice $x = 2a$, $g_2(x) \leq 0$. Hence $x = 2a$ and $\lambda = (2,0)$ is a valid KKT point.
- If $\lambda_2 \neq 0$ and $\lambda_1 \neq 0$. Then $x \in Y$ and $x = \frac{3}{2}a + r$. Then (4.5) yields

$$\frac{3}{2}a + r - \lambda_1(\frac{1}{2}a + r) + \lambda_2(-\frac{1}{2}a + r) = 0$$

  Which turns into

$$\begin{cases} 1 - \lambda_1 + \lambda_2 & = 0 \\ 3 - \lambda_1 - \lambda_2 & = 0 \end{cases}$$

  This implies $\lambda = (2,1)$. Any point in Y associated to $\lambda = (1,2)$ is a valid KKT point.

- **Step 4 : Second order info** We compute $H_x[\mathcal{L}](x) = (1 - \lambda_1 + \lambda_2)Id$
  - At the point $M = (x, \lambda)$, $x = 2a$ and $\lambda = (2, 0)$, we have $d \in V_M$ iff $(d|\nabla g_1(2a)) = 0$, that is $(d| - a) = 0$, the set V_M is not reduced to $\{0\}$. The Hessian is equal to $-Id$ so that the second order necessary condition can't be true and $x = 2a$ is not a local minimum.
  - At a point $M = (x, \lambda)$, $x = \frac{3}{2}a + r$ and $\lambda = (2, 1)$, the Hessian is equal to 0, the sufficient second order condition holds, only if V_M is $\{0\}$. The set V_M is the set of directions d such that

    $$(d|\frac{1}{2}a + r) = 0 \text{ and } (d| - \frac{1}{2}a + r) = 0$$

    Then the set V_M is the set of directions that are orthogonal to both a and r. It is reduced to $\{0\}$ only in dimension 2. Hence in dimension 2, we know that the sufficient conditions are verified and that these points are global minimizers. In dimension greater than 2, we cannot conclude yet, and we resort to computing the value of f at these point and we find that they all yield the same value. We can then conclude that these points are global minimizers.

  Hence Y is the set of global minimizers of f, there is no other local minimizers.

4.3 Finding maximizers

The goal is to establish the first and second order conditions for maximization problems. We first give the results and then proceed to the proof.

Proposition 4.3.1 Let f, and g_i be differentiable functions. We aim at solving $\sup_{x \in X} f(x)$, where

$$X = \{x \text{ such that } g_i(x) \leq 0 \text{ forall } i \in \mathcal{I} \text{ and } g_i(x) = 0 \text{ forall } i \in \mathcal{E}$$

- The Lagrangian $\mathcal{L}$ is defined by $\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x)$, the set Λ is defined by

  $$\Lambda = \{\lambda \in \mathbb{R}^p \text{ such that } \boldsymbol{\lambda_i \leq 0}, \ \forall i \in \mathcal{I}\}$$

- A point $M = (x, \lambda)$ is said to be a KKT point if

  $$\nabla_x \mathcal{L}(M) = 0, \quad x \in X, \quad \lambda \in \Lambda \text{ and } g_i(x)\lambda_i = 0 \quad \forall i$$

- At each point $M = (x, \lambda)$, the linearising cone of constraints is defined by

  $$V_M = \left\{ d \in \mathbb{R}^n \text{s.t.} \begin{cases} \langle \nabla g_i(x), d \rangle = 0 & \forall i \in \mathcal{E} \\ \langle \nabla g_i(x), d \rangle \leq 0 & \forall i \in \mathcal{A}_x \\ \lambda_i \langle \nabla g_i(x), d \rangle = 0 & \forall i \end{cases} \right\}$$

1. If $x^\star \in X$ is a local **maximizer** of f on X and the constraints are qualified at $x^\star$, then there exists $\lambda^\star$ such that $M^\star = (x^\star, \lambda^\star)$ is a KKT point. If in addition LICQ conditions holds at $x^\star \in X$, then

   $$\forall d \in V_{M^\star} \quad \langle H_x[\mathcal{L}](M^\star)d, d \rangle \leq 0.$$

2. If there exists $\lambda^\star$ such that $M^\star = (x^\star, \lambda^\star)$ is a KKT point and if
$$\forall d \in V_{M^\star} \text{ with } d \neq 0 \text{ then } \quad \langle H_x[\mathcal{L}](M^\star)d, d\rangle < 0,$$
then $x^\star$ is a local **maximizer** of f over X.

An important remark about maximization, is that the only thing that changes is the sign of λ_i for $i \in \mathcal{I}$ and the sign of the Hessian. Hence, when computing KKT point, it is very common to be able to solve the maximization and the minimization problem in one pass.

The proof is very simple, define $\tilde{f} = -f$ and $\tilde{\mathcal{L}}$ the associated Lagrangien and apply the first and second order conditions to $\tilde{\mathcal{L}}$. Note that

$$\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x) = -\left(-f(x) + \sum_i (-\lambda_i) g_i(x)\right) = -\tilde{\mathcal{L}}(x, -\lambda)$$

The KKT condition for the minimization problem is

$$\nabla_x \tilde{\mathcal{L}}(x, \lambda) = 0, \quad x \in X, \quad \lambda_i \geq 0 \ \forall i \in \mathcal{I} \text{ and } g_i(x)\lambda_i = 0 \quad \forall i$$

Replace λ by $-\lambda$ and then $\nabla_x \tilde{\mathcal{L}}(x, -\lambda)$ by $-\nabla_x \mathcal{L}(x, \lambda)$ to obtain :

$$\nabla_x \mathcal{L}(x, \lambda) = 0, \quad x \in X, \quad \lambda_i \leq 0 \ \forall i \in \mathcal{I} \text{ and } g_i(x)\lambda_i = 0 \quad \forall i$$

The definition of the linearized cone of constraints is unchanged and the Hessian is changed from $H_x[\tilde{\mathcal{L}}](x, -\lambda)$ to $-H_x[\mathcal{L}](x, \lambda)$, hence the change of direction of the inequality.

Chapter 5

Duality

In this chapter, we will focus on an optimization problem in the standard form :

$$
\begin{array}{lll}
(P) & \min\limits_{x\in\mathbb{R}^n} & f(x) \\
& \text{s.c.:} & g_i(x) \leq 0, i \in \mathcal{I} \ , \\
& & g_i(x) = 0, i \in \mathcal{E}
\end{array}
$$

where the functions f and g_i are C^1 and real-valued. According to KKT theorem, if the constraints are qualified, any solution $x^\star$ to such a problem can be associated to Lagrange multiplier $\lambda^\star \in \Lambda = \{\lambda \text{ such that } \lambda_i \geq 0 \forall i \in \mathcal{I}\}$. There exists two important families of algorithms that aim at solving the problem: the primal methods and the dual methods. The primal methods aim at finding a point $x^\star$, the Lagrange multiplier $\lambda^\star$ are just here to check that the point $x^\star$ is a KKT point. The primal methods work with a sequence of solutions, in most cases they deal with feasible solutions and a decrease of the objective function.

- Pros : When the algorithm stops, we obtain a **feasible** approximation of the solution.
- Cons: Theoritically and numerically difficult to obtain convergence.

On the other hand, the dual methods aim at finding the multipliers $\lambda^\star$ and not the point $x^\star$. This point can be deduced from the multipliers by **duality**.

- Pros: Robust methods (compared to primal methods), global convergence is easier to obtain
- Cons: A solution $x^\star$ can be only computed when the algorithm converged.

5.1 Inf-Sup duality

Consider the very general setting:

$$(P) \quad \inf_{x\in X} f(x),$$

where X is any subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the function to be minimized. The min-max duality pops in when f can be written as a supremum, that is:

$$f(x) = \sup_{y\in Y} \ \varphi(x, y)$$

where Y is a any set and $\varphi : X \times Y \to \mathbb{R}$ is a **coupling** function. The original problem is coined the **primal problem** $(\mathcal{P})$ and can be written as :

$$\text{PRIMAL PROBLEM :} \qquad \inf_{x\in X} \sup_{y\in Y} \varphi(x,y) \tag{$\mathcal{P}$}$$

The so-called **dual problem** is obtained by inverting the inf and the sup, it is given by

$$\text{DUAL PROBLEM :} \qquad \sup_{y\in Y} \inf_{x\in X} \varphi(x,y) \tag{$\mathcal{D}$}$$

We define the dual function of f in the following manner:

$$f^\star(y) = \inf_{x\in X} \ \varphi(x,y).$$

So that the dual problem is the one of maximization of $f^\star$ over Y. The only question that remains is : does solving the dual problem yields any information about the primal problem ? The first answer to this question is the weak dual principle

Theorem 5.1.1 — Weak duality. Let X, Y be non-empty sets and $\varphi : X \times Y \to \mathbb{R} \cup +\infty$. Then

$$\sup_{y\in Y} \inf_{x\in X} \ \varphi(x,y) \leqslant \inf_{x\in X} \sup_{y\in Y} \ \varphi(x,y).$$

The non-negative quantity $\inf_{x\in X} \sup_{y\in Y} \ \varphi(x,y) - \sup_{y\in Y} \inf_{x\in X} \ \varphi(x,y)$ is called the **duality gap**.

Proof

Recall that $f^\star(y) = \inf_{\tilde{x}\in X} \varphi(\tilde{x},y)$ and $f(x) = \sup_{\tilde{y}\in Y} \varphi(x,\tilde{y})$. We have, for all y and x

$$\inf_{\tilde{x}\in X} \varphi(\tilde{x},y) \leq \varphi(x,y) \leq \sup_{\tilde{y}\in Y} \varphi(x,\tilde{y})$$

So that $f^\star(y) \leq f(x)$, for every x and y. We can take the supremum in y and the infimum in x to obtain the weak duality theorem.

The key element that simplifies the analysis of min-max duality is the notion of **saddle point**

Definition 5.1.1 — Saddle-point. Let $\bar{x} \in X$ and $\bar{y} \in Y$. The point $(\bar{x}, \bar{y})$ is said to be a saddle point of φ on $X \times Y$ if :

$$\forall y \in Y \quad \varphi(\bar{x},y) \leq \varphi(\bar{x},\bar{y}) \leq \varphi(x,\bar{y}) \quad \forall x \in X$$

The saddle point allows us to caracterize when the dual problem and the primal problem are the same.

Theorem 5.1.2 — Strong duality. The point $(\bar{x}, \bar{y})$ is a saddle point of φ on $X \times Y$ iff

i. $\bar{x}$ is a solution of the primal problem $(\mathcal{P})$.
ii. $\bar{y}$ is a solution of the dual problem $(\mathcal{D})$.
iii. There is no duality gap, that is:

$$\sup_{y\in Y} \inf_{x\in X} \ \varphi(x,y) = \inf_{x\in X} \sup_{y\in Y} \ \varphi(x,y).$$

Proof

- Let $(\bar{x}, \bar{y})$ be a saddle point of φ and denote $\varphi^\star = \varphi(\bar{x}, \bar{y})$. For any x, we have $f(x) \geq \varphi(x, \bar{y}) \geq \varphi^\star$. But $f(\bar{x}) = \sup_y \varphi(\bar{x}, y) = \varphi(\bar{x}, \bar{y}) = \varphi^\star$. Hence $\inf f(x) = \varphi^\star$ and a solution of the primal problem is given by $\bar{x}$. By exactly the same argument $\sup f^\star(y) = \varphi^\star$ and a solution of the dual problem is given by $\bar{y}$. Note that there is no duality gap.
- Suppose now that $\bar{x}$ is a solution of the primal problem, $\bar{y}$ is a solution of the dual problem and that there is no duality gap, then denote $\varphi^\star = f(\bar{x}) = f^\star(\bar{y})$, we have
$$\forall y \in Y, \varphi(\bar{x}, y) \leq f(\bar{x}) = \varphi^\star.$$
Similarly, we have
$$\forall x \in X, \varphi(x, \bar{y}) \geq f^\star(\bar{y}) = \varphi^\star.$$
From the two equations above, we deduce $\varphi^\star = \varphi(\bar{x}, \bar{y})$, so that $(\bar{x}, \bar{y})$ is a saddle point.

5.2 Standard form and duality

From now on, we restrict our attention to problems in the standard form and we discuss about the dual problem of a problem in standard form

Definition 5.2.1 Consider the problem $\inf_{x \in X} f(x)$ where
$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i \in \mathcal{I},\ g_i(x) = 0,\ i \in \mathcal{E}\}.$$
Introduce the Lagrangian $\mathcal{L}$ defined on $\Lambda = \{\lambda,\ \lambda_i \geq 0 \text{ for all } \ i \in \mathcal{I}\}$ by
$$\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x).$$
Define the **primal function** $\bar{f}(x) = \sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda)$ then
$$\bar{f}(x) = \begin{cases} f(x) & \text{if } x \in X \\ +\infty & \text{if } x \notin X \end{cases}$$
So that the problem is equivalent to
$$\inf_{x \in \mathbb{R}^n} \bar{f}(x) = \inf_{x \in \mathbb{R}^n} \sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda).$$

Proof

The only thing to do is to compute $\bar{f}(x)$.

- If $x \in X$, then $g_i(x) \leq 0$ for each $i \in \mathcal{I}$ and then
$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i \in \mathcal{I}} \lambda_i g_i(x) \leq f(x).$$
Moreover $f(x) = \mathcal{L}(x, 0)$, hence $\sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda) = f(x)$.
- If $x \notin X$, then either there exists an $j \in \mathcal{I}$ such that $g_j(x) > 0$ or there exists an $j \in \mathcal{E}$ such that $g_j(x) \neq 0$.

- If there exists an $j \in \mathcal{I}$ such that $g_j(x) > 0$, take λ such that $\lambda_i = 0$ except for $i = j$, and then $\mathcal{L}(x, \lambda) = f(x) + \lambda_j g_j(x)$. Let λ_j go to $+\infty$ and then $\sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda) = +\infty$.
- If there exists an $j \in \mathcal{E}$ such that $g_j(x) \neq 0$, take λ such that $\lambda_i = 0$ except for $i = j$, and then $\mathcal{L}(x, \lambda) = f(x) + \lambda_j g_j(x)$. Let λ_j go to $+\infty$ or $-\infty$, depending on the sign of $g_j(x)$ and then $\sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda) = +\infty$.

Following the discussion of the previous section, the dual problem is obtained by exchanging the inf and the sup, we define

- The **dual function** is given by

$$f^\star(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda).$$

- The **dual problem** is given by:

$$\sup_{\lambda \in \Lambda} f^\star(\lambda)$$

Proposition 5.2.1 The dual function $f^\star$ is always concave and Λ is a convex set.

Proof

The set Λ is clearly convex. Remark that the function $\lambda \mapsto \mathcal{L}(x, \lambda)$ is affine and hence concave. The function $f^\star$ is the infimum of concave functions and such an infimum is always concave. This is a consequence of the following proposition (which proof follows) : "*Let I be any set of indices, and f_i be convex functions for each $i \in I$, then $f = \sup_{i \in I} f_i$ is a convex function*". In order to prove this assertion, for any function $f : A \to \mathbb{R}$, define $Epi(f)$ the epigraph of f as the set

$$Epi(f) = \{(x, t) \in A \times \mathbb{R} \text{ such that } f(x) \leq t\}.$$

Then prove three easy lemmatas

- The function g is convex iff $Epi(g)$ is a convex set.
- For any real valued function if $f = \sup_{i \in I} f_i$, then $Epi(f) = \cap_{i \in I} Epi(f_i)$
- For any set of indices if $\mathcal{C}_i$ is a convex set for all $i \in I$, then $\cap_{i \in I} \mathcal{C}_i$ is a convex set.

We already know that if $(\bar{x}, \bar{\lambda})$ is a saddle-point of the Lagrangian, then $\bar{x}$ is a solution of the primal problem. An important feature of saddle point is that they verify KKT equations, without the need for qualification of constraints.

Theorem 5.2.2 — Saddle points are KKT. Consider the problem $\inf_{x \in X} f(x)$ where

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, \ i \in \mathcal{I}, \ g_i(x) = 0, \ i \in \mathcal{E}\},$$

where f and g_i are C^1 functions for each i. Let $\Lambda = \{\lambda, \ \lambda_i \geq 0, \ i \in \mathcal{I}\}$. If a point $(\bar{x}, \bar{\lambda}) \in X \times \Lambda$ is a saddle point of the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x).$$

Then it verifies KKT conditions that is $\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = 0$ and $\bar{\lambda}_i g_i(\bar{x}) = 0$ for all i.

Proof

Let $(\bar{x}, \bar{\lambda})$ be a saddle-point of the Lagrangian. Then

$$\mathcal{L}(\bar{x}, \bar{\lambda}) \leq \mathcal{L}(x, \bar{\lambda}) \quad \forall x.$$

So that $\bar{x}$ is a minimum of the unconstrained function $x \mapsto \mathcal{L}(x, \bar{\lambda})$. Hence, we must have

$$\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = 0$$

Because $\bar{x} \in X$, we have $g_i(\bar{x}) = 0$ for every $i \in \mathcal{E}$ and $\mathcal{L}(\bar{x}, \lambda) = f(\bar{x}) + \sum_{i \in \mathcal{I}} \lambda_i g_i(\bar{x})$ Recall that we have

$$\mathcal{L}(\bar{x}, \lambda) \leq \mathcal{L}(\bar{x}, \bar{\lambda}) \quad \forall \lambda \in \Lambda$$

For the particular choice $\lambda = 0$, this yields $\sum_{i \in \mathcal{I}} \bar{\lambda}_i g_i(\bar{x}) \geq 0$. Since $\bar{\lambda}_i \geq 0$ and $g_i(\bar{x}) \leq 0$, this implies $\bar{\lambda}_i g_i(\bar{x}) = 0$ for all $i \in \mathcal{I}$.

Hence saddle points are KKT points, the converse is true when the problem is convex.

Theorem 5.2.3 — Convex Lagrangian duality. Consider the problem $\inf_{x \in X} f(x)$ where

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i \in \mathcal{I},\ g_i(x) = 0,\ i \in \mathcal{E}\}.$$

where the functions g_i and f are C^1 and convex and the functions g_i, $i \in \mathcal{E}$ are affine. Let $\Lambda = \{\lambda,\ \lambda_i \geq 0,\ i \in \mathcal{I}\}$. A point $(\bar{x}, \bar{\lambda}) \in X \times \Lambda$ is a saddle point of the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x)$$

if and only if it verifies KKT conditions that is $\nabla_x \mathcal{L}(x, \bar{\lambda}) = 0$ and $\bar{\lambda}_i g_i(\bar{x}) = 0$ for all i. As a corollary, if $\bar{x}$ is a local minimum of the primal problem and if the constraints are qualified at $\bar{x}$, then there exists $\bar{\lambda}$ such that $(\bar{x}, \bar{\lambda})$ is a saddle point.

Proof

We already know that saddle points verify the KKT conditions. Let $(\bar{x}, \bar{\lambda})$ be a KKT point of $(\mathcal{P})$, then $\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = 0$. Hence $\bar{x}$ is a critical point of

$$\psi : x \in \mathbb{R}^n \mapsto \mathcal{L}(x, \bar{\lambda}) = f(x) + \sum_{i \in \mathcal{I}} \bar{\lambda}_i g_i(x) + \sum_{i \in \mathcal{E},\ \bar{\lambda}_i \geq 0} \bar{\lambda}_i g_i(x) + \sum_{i \in \mathcal{E},\ \bar{\lambda}_i \leq 0} (-\bar{\lambda}_i)(-g_i(x)).$$

But ψ is a positive linear combination of convex function hence is convex, and $\bar{x}$ is a global minimum of ψ and we have

$$\mathcal{L}(\bar{x}, \bar{\lambda}) \leq \mathcal{L}(x, \bar{\lambda}) \quad \forall x$$

On the other hand, for every $\lambda \in \Lambda$, we have

$$\mathcal{L}(\bar{x}, \lambda) = f(\bar{x}) + \sum_{i \in \mathcal{I}} \lambda_i g_i(\bar{x}) \leq f(\bar{x}) = \mathcal{L}(\bar{x}, \bar{\lambda}).$$

Hence $(\bar{x}, \bar{\lambda})$ is a saddle point of the Lagrangian.

5.3 Duality of Linear Programming

In Linear Programming (LP in short), the definition of duality is a little bit different. Indeed, LP duality is tailored such that the definition of the dual problem is more "symmetric". Understanding LP duality allows the student to kill two birds with the same stone. First LP-problems are very common problems with very strong properties. Second, there may be slightly different definitions of duality that leads to slightly different problems.

A Linear Programming problem is a problem of the form $\inf_{x \in X} \langle c, x \rangle$ where X is defined via affine functions only. There are two kind of constraints, the one that are stated on the variable x only, they are of the form $x_i \geq 0$, $x_i \leq 0$ or $x_i = 0$ and the one that are of the form $\langle a_j, x \rangle \leq b_j$ or $\langle a_j, x \rangle = b_j$ or $\langle a_j, x \rangle \geq b_j$. We build the matrix A whose rows are exactly the vectors a_j, we can then split the set of constraints into two sets $X = X_a \cap X_b$, the set X_a that has only constraints of the form

$$X_a = \left\{ x \text{ s.t. } \begin{cases} x_i \leq 0 & \text{if } i \in \mathcal{A}^- \\ x_i \in \mathbb{R} & \text{if } i \in \mathcal{A} \\ x_i \geq 0 & \text{if } i \in \mathcal{A}^+ \end{cases} \right\}, X_b = \left\{ x \text{ s.t. } \begin{cases} (Ax)_j \leq b_j & \text{if } j \in \mathcal{B}^- \\ (Ax)_j = b_j & \text{if } j \in \mathcal{B} \\ (Ax)_j \geq b_j & \text{if } j \in \mathcal{B}^+ \end{cases} \right\}$$

There are several standard tricks that allows any Linear Programming problem to be put on one of the two following form

Proposition 5.3.1 Each linear programming problem can be put in one of the two following forms

- **Inequality form :** $\inf\limits_{\hat{A}\hat{x} \leq \hat{b}} \langle \hat{x}, \hat{c} \rangle$.
- **Equality form :** $\inf\limits_{\tilde{x} \geq 0, \tilde{A}\tilde{x} = \tilde{b}} \langle \tilde{x}, \tilde{c} \rangle$.

Proof

We first show how to put a standard problem in inequality form and then how to put it in equality form, there are other ways to proceed, but we are only interested in showing that there exists a transformation.

- Put a problem in inequality form
  - **Get rid of X_a:** : Take any problem and note that the constraints on X_a can be transfered on constraints in X_b. Indeed, a constraint $x_i \geq 0$ is just a constraint of the form $\langle e_i, x \rangle \geq 0$. So up to adding e_i as the last row of A and 0 as the last element of b, any constraint in X_a with $i \in \mathcal{A}^-$ (resp. $i \in \mathcal{A}^+$) can be transfered to a constraint in X_b with $i \in \mathcal{B}^-$ (resp. $i \in \mathcal{B}^+$).
  - **Get rid of $\mathcal{B}$:** If there is an equation of the form $(Ax)_j = b_j$, transform it into two inequalities $(Ax)_j \leq b_j$ and $(Ax)_j \geq b_j$. This means that the j^{th} row of A and the j^{th} element of b are copied, this transforms an equality constraint into two inequality constraints
  - **Get rid of $\mathcal{B}^+$:** For any equation of the form $(Ax)_j \geq b_j$, multiply it by -1 and transform it into $-(Ax)_j \leq -b_j$. This multiplies the j^{th} row of A and the j^{th} coefficient of b by -1.
- Transform problem from inequality form to equality form.
  - **Transform $\mathcal{A}$ into $\mathcal{A}^+$:** We first want to transform a problem of the form $X = \{Ax \leq b\}$ into a problem of the form $X = \{\hat{x} \geq 0, \hat{A}\hat{x} \leq \hat{b}\}$, hence we want to transform constraints of the form $\mathcal{A}$ into constraints of the form

$\mathcal{A}^+$. In order to do so, use the fact that any real number can be written (in a non unique way) as the difference of two positive numbers. For each $i \in \mathcal{A}$, denote $x_i^+ \geq 0$ and $x_i^- \geq 0$ such that $x_i = x_i^+ - x_i^-$. Denote

$$\hat{x} = \begin{pmatrix} x^+ \\ x^- \end{pmatrix}, \quad \hat{A} = \begin{pmatrix} A & -A \end{pmatrix} \quad \hat{b} = b. \quad \hat{c} = \begin{pmatrix} c \\ -c \end{pmatrix}.$$

Then the problems $\inf_{Ax \leq b} \langle c, x\rangle$ is equivalent to $\inf_{\hat{x} \geq 0, \hat{A}\hat{x} \leq \hat{b}} \langle \hat{c}, \hat{x}\rangle$.

- **Transform $\mathcal{B}^-$ into $\mathcal{B}$:** Suppose that one is given a problem of the form $\inf_{x \geq 0, Ax \leq b} \langle c, x\rangle$. The idea is to transform any constraint $(Ax)_j \leq b_j$ into $(Ax)_j + \lambda_j = b_j$ with $\lambda_j \geq 0$. In order to do so, construct the unknown vector $\hat{x}$ and the matrix $\hat{A}$ the following way

$$\hat{x} = \begin{pmatrix} x \\ \lambda \end{pmatrix} \quad \hat{A} = \begin{pmatrix} A & Id \end{pmatrix}, \quad \hat{b} = b, \quad \hat{c} = \begin{pmatrix} c \\ 0 \end{pmatrix}.$$

The constraint $\hat{A}\hat{x} = b$ and $\lambda \geq 0$ is then exactly $Ax \leq b$. Note that since $x \geq 0$, then the problem is exactly in equality form.

Before talking about the duality of a linear programming, we note that there exists solutions to a linear program unless the solutions lies at infinity or if there is no admissible points.

Theorem 5.3.2 There exists a solution if and only if the optimal value is finite, that is

$$-\infty < \inf_{x \in X} \langle c, x\rangle < +\infty \iff \exists x^\star \in X \text{ s.t. } \langle c, x^\star\rangle = \inf_{x \in X} \langle c, x\rangle$$

Note that $\inf_{x \in X} \langle c, x\rangle = +\infty$ if and only if the problem has no admissible point (X is empty).

Proof

Suppose that the optimal value is finite. Put the problem in $\inf\limits_{x \geq 0, Ax = b} \langle c, x\rangle$. Take x_k a minimizing sequence so that $\langle c, x_k\rangle$ converge to some $\alpha \in \mathbb{R}$ and introduce the matrix

$$\mathcal{A} = \begin{pmatrix} c^T \\ A \end{pmatrix}.$$

Denote $(\mathcal{A})_i$ the columns of $\mathcal{A}$ and let $z_k = \mathcal{A}x_k = \sum_i (x_k)_i \mathcal{A}_i$. For every k, z_k belongs to the cone

$$\mathcal{C} = \left\{ \sum_i \lambda_i \mathcal{A}_i \text{ with } \lambda_i \geq 0 \right\}.$$

Moreover $z_k = \begin{pmatrix} \langle c, x_k\rangle \\ b \end{pmatrix}$ converges to $\begin{pmatrix} \alpha \\ b \end{pmatrix}$. From the proof of Theorem 3.2.1, we know that $\mathcal{C}$ is closed, there exists some $\lambda \geq 0$ such that $\mathcal{A}\lambda = \begin{pmatrix} \alpha \\ b \end{pmatrix}$. The minimum is then attained at λ.

The special thing about Lagrangian duality for linear programming is that not all the constraints are put inside the Lagrangian. The theorem reads as follows :

Theorem 5.3.3 Suppose that A is a $m\times n$ matrix, x and c are vectors of $\mathbb{R}^n$, and y and b are vectors of $\mathbb{R}^m$. Suppose that $\mathcal{A}, \mathcal{A}^+, \mathcal{A}^-$ is a partition of $[1,n]$ and $\mathcal{B}, \mathcal{B}^+, \mathcal{B}^-$ is a partition of $[1,m]$ and

$$X_a = \left\{ x \text{ s.t. } \begin{cases} x_i \le 0 & \text{if } i \in \mathcal{A}^- \\ x_i \in \mathbb{R} & \text{if } i \in \mathcal{A} \\ x_i \ge 0 & \text{if } i \in \mathcal{A}^+ \end{cases} \right\}, \qquad X_b = \left\{ x \text{ s.t. } \begin{cases} (Ax)_j \le b_j & \text{if } j \in \mathcal{B}^- \\ (Ax)_j = b_j & \text{if } j \in \mathcal{B} \\ (Ax)_j \ge b_j & \text{if } j \in \mathcal{B}^+ \end{cases} \right\}$$

$$Y_a = \left\{ y \text{ s.t. } \begin{cases} (A^T y)_i \ge c_i & \text{if } i \in \mathcal{A}^- \\ (A^T y)_i = c_i & \text{if } i \in \mathcal{A} \\ (A^T y)_i \le c_i & \text{if } i \in \mathcal{A}^+ \end{cases} \right\}, \quad Y_b = \left\{ y \text{ s.t. } \begin{cases} y_j \le 0 & \text{if } j \in \mathcal{B}^- \\ y_j \in \mathbb{R} & \text{if } j \in \mathcal{B} \\ y_j \ge 0 & \text{if } j \in \mathcal{B}^+ \end{cases} \right\}$$

Introduce the Lagrangian

$$\mathcal{L}(x,y) = \langle c, x\rangle + \langle y, b\rangle - \langle y, Ax\rangle$$

Then

$$\sup_{y\in Y_b} \mathcal{L}(x,y) = \begin{cases} \langle c, x\rangle & \text{if } x \in X_b \\ +\infty & \text{if not} \end{cases} \quad \text{and} \quad \inf_{x\in X_a} \mathcal{L}(x,y) = \begin{cases} \langle b, y\rangle & \text{if } y \in Y_a \\ -\infty & \text{if not} \end{cases}$$

So that the following two problems are in duality

$$\begin{aligned} \inf_{x\in X_a\cap X_b} \langle c, x\rangle &= \inf_{x\in X_a} \sup_{y\in Y_b} \mathcal{L}(x,y) \\ \sup_{y\in Y_a\cap Y_b} \langle b, y\rangle &= \sup_{y\in Y_b} \inf_{x\in X_a} \mathcal{L}(x,y) \end{aligned}$$

The rule to construct the dual problem for the primal is the following : denote a_i the rows of A and $a_j^\star$ the columns of A.

Minimize $\langle x, c\rangle$	$\Longleftrightarrow$	Maximize $\langle y, b\rangle$
$\langle a_j, x\rangle \ge b_j$		$y_j \ge 0$
$\langle a_j, x\rangle = b_j$	$\Longleftrightarrow$	$y_j \in \mathbb{R}$
$\langle a_j, x\rangle \le b_j$		$y_j \le 0$
$x_i \ge 0$		$\langle a_i^\star, y\rangle \le c_i$
$x_i \in \mathbb{R}$	$\Longleftrightarrow$	$\langle a_i^\star, y\rangle = c_i$
$x_i \le 0$		$\langle a_i^\star, y\rangle \ge c_i$

Theorem 5.3.4 Consider a Linear Programming problem and its duality as stated in Theorem 5.3.3.

- If one of the problems (primal or dual) has a solution then they both admit a solution $(\bar{x}, \bar{y})$ which is a saddle point of the Lagrangian.
- If (x,y) is a couple of admissible points of the primal and dual problems, then
$$\langle c, x\rangle \ge \langle y, b\rangle. \tag{5.1}$$
- If both problems (primal or dual) have non-empty admissible points, then hey both admit a solution $(\bar{x}, \bar{y})$ which is a saddle point of the Lagrangian.
- If (x,y) is a couple of admissible points of the primal and dual problems, then $\langle c, x\rangle = \langle y, b\rangle$ if and only if (x,y) is a saddle point of the Lagrangian.

Proof

- All the constraints are affine, the constraints are qualified so that a solution of the primal (or the dual) is a KKT point. Then a linear programming problem is convex, so that saddle points are exactly KKT points.
- By definition, for each $x \in X_a$ and $y \in Y_a$, we have, for all i $x_i c_i \geq (A^T y)_i x_i$. We conclude that $\langle c, x\rangle \geq \langle A^T y, x\rangle$. Similarly, for each $x \in X_b$ and $y \in Y_b$, we have, for all j $y_j (Ax)_j \geq y_j b_j$. We conclude that $\langle y, Ax\rangle \geq \langle y, b\rangle$. We conclude then that
$$\langle c, x\rangle \geq \langle A^T y, x\rangle = \langle y, Ax\rangle \geq \langle y, b\rangle$$
- If both problems have non-empty admissible points, then by $\langle c, x\rangle \geq \langle y, b\rangle$, the primal problem is lower-bounded and hence admits solutions. Similarly for the dual problem.
- Suppose that $(\bar{x}, \bar{y})$ is admissible and that
$$\langle c, \bar{x}\rangle = \langle \bar{y}, b\rangle.$$
For all y admissible, (5.1) is true if x is replaced by $\bar{x}$. Hence
$$\langle \bar{y}, b\rangle = \langle c, \bar{x}\rangle \geq \langle y, b\rangle.$$
Which proves that $\bar{y}$ is a solution to the primal problem and similarly we can show that $\bar{x}$ is a solution to the dual problem. Equality in (5.1) means that the duality gap is zero. This means that $(\bar{x}, \bar{y})$ is a saddle-point. In the other hand, if we have a saddle-point, we must have zero duality-gap, hence equality in (5.1).

II

Algorithmics of smooth optimization

CHAPTER 6

First order descent methods

6.1 Fonctions with Lipschitz gradient

A Lipschitz function is a continuous function which is almost differentiable, in the sense that its rate of increase is bounded, ie. we say that $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz at the point x if there exists $C > 0$ such that for every h, we have :

$$\frac{|f(x+h) - f(x)|}{h} \leq C.$$

In the Taylor expansions, we have the following caracterization of the expansions

- Continuity : $f(x+h) = f(x) + o(1)$
- Lipschitz : $f(x+h) = f(x) + \mathcal{O}(h)$
- Derivable : $f(x+h) = f(x) + f'(x)h + o(h)$
- Lipschitz-derivative : $f(x+h) = f(x) + f'(x)h + \mathcal{O}\big(h^2\big)$

It turns out that a very important class of functions for optimization is the class of differentiable functions with Lipschitz gradient. It is the class of functions who verify

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \mathcal{O}\big(\|h\|^2\big)\,.$$

The constant in the remainder can be made explicit, as in the following proposition

Proposition 6.1.1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **differentiable with a L-Lipschitz gradient.** if it is differentiable and if there exists $L \geq 0$ such that, forall x and $y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{6.1}$$

Moreover, we have

$$\left| f(y) \underbrace{- f(x) - \langle \nabla f(x), y - x \rangle}_{\text{Taylor expansion}} \right| \leq \frac{L}{2}\|y - x\|^2.$$

Proof

First denote $h = y - x$ and $\phi : t \mapsto f(x+th)$, we can verify that $\phi'(t) = \langle \nabla f(x+th), h \rangle$ and from

$$\phi(1) = \phi(0) + \int_0^1 \phi'(t)dt,$$

we obtain:

$$\begin{aligned} f(x+h) &= f(x) + \int_0^1 \langle \nabla f(x+th), h \rangle dt \\ &= f(x) + \int_0^1 \langle \nabla f(x+th) - \nabla f(x) + \nabla f(x), h \rangle dt \\ &= f(x) + \langle \nabla f(x), h \rangle + \underbrace{\int_0^1 \langle \nabla f(x+th) - \nabla f(x), h \rangle dt}_{=\mathbf{(A)}} \end{aligned}$$

The term $\mathbf{(A)}$ can be bounded using (6.1) and the Cauchy-Schwartz inequality, indeed we have:

$$|\langle \nabla f(x+th) - \nabla f(x), h \rangle| \quad \leq \quad \|\nabla f(x+th) - \nabla f(x)\| \|h\| \leq Lt\|h\|^2$$

we obtain

$$\begin{aligned} |f(x+h) - f(x) - \langle \nabla f(x), h \rangle| \quad &= \quad |\mathbf{(A)}| \leq \int_0^1 tL\|h\|^2 dt \\ &\leq \quad \frac{L}{2}\|h\|^2. \end{aligned}$$

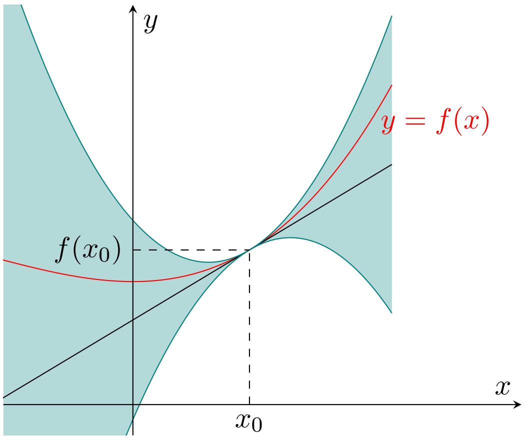


Figure 6.1: An example of function with Lipschitz gradient, we represent the function f in red, its tangent and the two parabolas $x \mapsto f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \pm \frac{L}{2}\|x - x_0\|^2$. The function f is allowed to take any value in-between the two parabolas. It cannot stray too far apart from the first order approximation.

If a function is C^2, then the spectrum of the Hessian yields the Lipschitz constant of the gradient.

Proposition 6.1.2 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2 function. For each x, denote $\mathrm{Sp}(H[f](x))$ the set of eigenvalues of the Hessian of f at point x. Then

$$L = \sup_{x\in\mathbb{R}^n} \sup_{\lambda\in \mathrm{Sp}(H[f](x))} |\lambda|.$$

Exercice 6.1

1. The function $f(x) = \frac{1}{2}\|Ax - b\|^2$ has Lipschitz gradient of constant $L = \lambda_{max}(A^TA)$.
2. The function $f(x) = -\log(x)$ has no Lipschitz gradient on $\mathbb{R}_+^*$. On each interval $[a, +\infty[$ with $a > 0$, then f has a a^{-2}-Lipschitz gradient.
3. The function $f(x) = \exp(x)$ has no Lipchitz gradient on $\mathbb{R}$. On each interval $]-\infty, b]$, then f has a $\exp(b)$-Lipschitz gradient.

In this chapter, we design numerical methods that aim at minimizing unconstrained optimization problems. The problem we solve is then written as

$$(P) \qquad \min_{x\in\mathbb{R}^n} f(x),$$

where f is a function from $\mathbb{R}^n$ to $\mathbb{R}$ which is differentiable. For some algorithm, we will even suppose that f is C^2. We focus on the so-called descent methods, these methods guarantee at each iteration that the function decreases

6.2 Basic notions of optimization algorithms

6.2.1 Description of descent methods

The descent methods are iterative algorithms that start at a point x_0 and generate a sequence of iterates $(x_k)_{k\in\mathbb{N}}$ such that

$$x_{k+1} = x_k + s_k d_k.$$

The iterates will verify the non-increasing condition

$$\forall k \in \mathbb{N}, \quad f(x_{k+1}) \leq f(x_k).$$

Three things determine the algorithm,

- The choice of d_k, which is called the **direction**.
- The choice of s_k, which is called the **step**. The decision algorithm that choses the step is called the **linear search**.
- The choice of **stopping criterion**.

A key notion for the choice of direction d_k is to pick up a **direction of descent**, in a nutshell a direction of descent is a direction in which the function decreases for small steps.

Definition 6.2.1 — Direction of descent. We say that a direction d is a **direction of descent of f at point x** if

$$\langle \nabla f(x), d\rangle < 0$$

If d is a direction of descent of f at point x, there exists $r > 0$ such that forall

$0 < s < r$

$$f(x+sd) < f(x)$$

Proof

We have

$$f(x+sd) = f(x) + s\langle \nabla f(x), d\rangle + o(s) = f(x) + s\left(\langle \nabla f(x), d\rangle + o(1)\right)$$

Since $\langle \nabla f(x), d\rangle < 0$ there exists $r > 0$ so that for all $0 < s < r$, we have

$$\langle \nabla f(x), d\rangle + o(1) < 0$$

and the proof is complete.

In other words, directions of descent ensure that there exists a threshold r, such that for each choice of step $0 < s_k < r$, then $f(x_{k+1}) \leq f(x_k)$. In other words, if d_k is a direction of descent, upon taking a small enough step, we are sure that the algorithm is a strict descent (the objective function is decreasing) algorithm.

```
STANDARD DESCENT METHOD.
 Input: $f : \mathbb{R}^n \to \mathbb{R}$ differentiable, $x_0$ arbitray intial point.
Output: an approximation of $\min_{x \in \mathbb{R}^n} f(x)$.
  • $k := 0$
  • While Convergence criterion is not met,
      • Descent direction: Find a direction $d_k$ such that $\langle \nabla f(x_k), d_k\rangle < 0$.
      • Line search: Choose a step $s_k > 0$ such that
            $f(x_k + s_k d_k) < f(x_k)$.
      • Update: $x_{k+1} = x_k + s_k d_k$; $k := k + 1$;
  • Return $x_k$.
```

6.2.2 Stopping criterion

First, remark that descent algorithm are stuck at critical point. Indeed if $\nabla f(x_k) = 0$, it is impossible to find a descent direction in the sense of Definition 6.2.1. The best we can hope is to converge to a local minimizer and not a global minimizer.

Second, we always need to bound *a priori* the number of iterations. This prevents the algorithm for running in a infinite time if it enters a loop. Indeed, note that even algorithm which are known to converge with a prescribed rate of convergence can loop if -for instance- numerical error is above the tolerance treshold.

Let $\varepsilon > 0$ be the asked precision. We have several criterion at our disposal. The first one is an optimality criterion based on necessary conditions of first order. In the case of unconstrained optimization, we will step if

$$\|\nabla f(x_k)\| < \varepsilon, \tag{6.2}$$

and the algorithm will return x_k as an approximation of the local minimizer.

In practice, the algorithm may fail to satisfy the test (6.2), it will surely be the case if -for example- the user sets ε to be greater than machine precision. We have several other tests at our disposal:

- Stagnation of the solution: $\|x_{k+1} - x_k\| < \varepsilon(1 + \|x_k\|)$.
- Stagnation of the objective: $\|f(x_{k+1}) - f(x_k)\| < \varepsilon(1 + |f(x_k)|)$.

We usually implement one of the two criterion above if the algorithm seems to stop converging. The recipe for a good stopping criterion is

$$\text{maximum number of iteration } + (6.2) + \text{ if needed, stagnation criteria}$$

In practice, one deals with relative errors and not absolute one. Moreover, some algorithm have very strong convergence result, if these convergence results leads to a usable criterion, one should favor such a criterion.

6.2.3 Speed of convergence

We first define the notion of convergence for an optimization algorithm. As it turns out there are several notions, some stronger than others.

Definition 6.2.2 — Convergence of an optimization algorithm. We say that an algorithm **converges to a critical point** if

$$\lim_{k \to +\infty} \|\nabla f(x_k)\| = 0.$$

We say that an algorithm **converges in value** if

$$\lim_{k \to +\infty} f(x_k) = \inf_{x \in X} f(x).$$

We say that the **iterates converges** if there exists $x^\star \in X$ such that

$$\lim_{k \to +\infty} x_k = x^\star.$$

If any of the above convergence is met only for a subsequence and not for the full sequence, the convergence is said to hold **up to a subsequence**.

Attention, the notion of **convergence to a critical point** or **convergence of the iterates** does not ensure that the algorithm converges towards a minimum, even a local minimum. Take for instance the function $f : (x, y) \mapsto x^2 - y^2 + y^4$ which is coercive. Its global minimum is attained for the points $M_\pm = (0, \pm 1/\sqrt{2})$. However, start at the point $M_0 = (1, 0)$, and take the following choice

$$d_k = (-2x_k, 2y_k - 3y_k^3), \quad s_k << 1.$$

Then this algorithm is a descent algorithm (we have $f(M_{k+1}) \leq f(M_k)$ and its iterates converges to $(0, 0)$ which is a critical point but $(0, 0)$ is not a global minimizer.

Definition 6.2.3 Let $(x_k)_{k \in \mathbb{N}}$ be a converging sequence towards $x^\star = \lim_k x_k$. We say that the convergence is

- **linear** if there exists $\tau \in]0, 1[$ such that:
$$\lim_{k \to +\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = \tau.$$
- **superlinear** if
$$\lim_{k \to +\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = 0.$$

- **of order p** if there exists $\tau \geq 0$ such that:

$$\lim_{k \to +\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|^p} = \tau.$$

6.3 Classical convergence of gradient algorithm

6.3.1 Description of the algorithm

Amongst the direction of descent, choosing the direction opposite to the gradient is a method of choice known as the **gradient method**.

Proposition 6.3.1 If $\nabla f(x_k) \neq 0$, then choosing $d_k = -\nabla f(x_k)$ is called the **gradient method**. It is a descent direction kown as the **steepest direction** for the following property: For any $m > 0$, any solution of the following problem :

$$\inf_{\|d_k\| \leq m} f(x_k) + \langle \nabla f(x_k), d_k \rangle$$

is a positive multiple of $-\nabla f(x_k)$.

Proof

We fix $m > 0$ and we denote $d^\star$ the solution of

$$\inf_{\|d\| \leq m} f(x_k) + \langle \nabla f(x_k), d \rangle.$$

We have to show that $d^\star$ exists, is unique and can be written as $d^\star = -\alpha \nabla f(x_k)$ for some $\alpha > 0$. Because $f(x_k)$ is a constant, we focus on

$$\inf_{\|d\| \leq m} \langle \nabla f(x_k), d \rangle.$$

There are two ways to prove this. The first way relies on Cauchy-Schwarz inequality, we always have

$$\langle \nabla f(x_k), d \rangle \geq -\|\nabla f(x_k)\| \|d\| \geq -m \|\nabla f(x_k)\|$$

With equality in the first inequality if and only if d and $\nabla f(x_k)$ are colinear and have opposite direction, that is $d = -\alpha \nabla f(x_k)$, $\alpha \geq 0$. The exact value of α is $\alpha = \frac{m}{\|\nabla f(x_k)\|}$ so that $\|d\| = m$ and there is equality in the second inequality. For the second proof, remark that the problem has a solution (continuous function on a bounded closed set), then remark that the constraint $g(d) = \|d\|^2 - m^2 \leq 0$ is qualified, indeed its gadient is

$$\nabla g(d) = 2d$$

is non-zero when $g(d) = 0$. We write down KKT equations which are

$$\begin{cases} \nabla f(x_k) + 2\lambda d = 0 \quad \lambda g(d) = 0 \end{cases}$$

The case $\lambda = 0$ is impossible so that setting $\alpha = \frac{1}{2\lambda} > 0$, we have $d = -\alpha \nabla f(x_k)$, α and m are related by the equation $\alpha = \frac{m}{\|\nabla f(x_k)\|}$.

Proposition 6.3.1 gives a nice interpretation of a gradient algorithm, it follows the discussion:

1. Given x_k, minimizing $\min_{x\in\mathbb{R}^n} f(x)$ amounts to finding $d^\star$ solution to $\min_{d\in\mathbb{R}^n} f(x_k+d)$ and to return $x_k + d^\star$.
2. We suppose that $d^\star$ is small, that is we managed to get close to the actual minimizer, we replace f by its first order Taylor expansion, the goal is to minimize
$$\inf_{d\in\mathbb{R}^n} f(x_k) + \langle \nabla f(x_k), d\rangle$$
3. Damn !! the above problem has no solution, indeed setting $d = -t\nabla f(x_k)$ with $t \to +\infty$ shows that the above inf is equal to $-\infty$. But this solution has no meaning because in this case d is very large, and we supposed that d was small !! Hence we enforce smallness of d by looking for solutions of the form
$$\inf_{\|d\|\leq m} f(x_k) + \langle \nabla f(x_k), d\rangle$$
4. Great, we find $x_{k+1} = x_k + s_k d_k$ with $d_k = -\nabla f(x_k)$ and α_k directly linked to the choice of m.

This shows that the gradient method can be interpreted as a method where the function is replaced at each iteration by its first order Taylor expansion.

6.3.2 Convergence of Gradient algorithm

Theorem 6.3.2 — Convergence of steepest descent algorithm. Let f be a C^1 function, bounded from below and with Lipschitz gradient of constant L. We consider the choice of direction of descent $d_k = -\nabla f(x_k)$ and $x_{k+1} = x_k + s_k d_k$

- If $s_k < \frac{2}{L}$, then $f(x_{k+1}) < f(x_k)$.
- If $s < \frac{2}{L}$, the fixed step algorithm $s_k = s$ converges. As a rule of thumb, the best choice is $s = \frac{1}{L}$
- The optimal step gradient algorithm converges.

by "convergence", we mean that $\sum_k \|\nabla f(x_k)\|^2 < +\infty$.

Proof We recall that $x_{k+1} = x_k - s_k\nabla f(x_k)$.

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) - s_k\|\nabla f(x_k)\|^2 + \frac{L}{2}s_k^2\|\nabla f(x_k)\|^2 \\
&\leq f(x_k) + s_k\left(\frac{L}{2}s_k - 1\right)\|\nabla f(x_k)\|^2. \qquad (6.3)
\end{aligned}$$

- If $s_k < \frac{2}{L}$, then $f(x_{k+1}) < f(x_k)$.
- For the fixed step algorithm with $s < \frac{2}{L}$, there exists a $c > 0$ such that
$$f(x_k) - f(x_{k+1}) \geq c\|\nabla f(x_k)\|^2.$$
Adding up those inequalities, we obtain
$$f(x_0) - f(x_{n+1}) \geq \sum_{k=0}^{n} c\|\nabla f(x_k)\|^2.$$

We use $f(x_{n+1}) \geq \inf_x f(x)$ to obtain

$$\sum_{k=0}^{+\infty} c\|\nabla f(x_k)\|^2 \leq f(x_0) - \inf_{x\in\mathbb{R}^n} f(x).$$

Moreover if we minimize the right-hand side of (6.3) with respect to s_k, we find that the best s_k is L^{-1}.

- For the optimal step case, since the step s_k minimizes $f(x_{k+1})$, then we must have

$$f(x_{k+1}) \leq f(x_k - L^{-1}\nabla f(x_k))$$

Use (6.3) with $s_k = L^{-1}$ to obtain :

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

and proceed as in the fixed step algorithm.

From the convergence of the sum $\|\nabla f(x_k)\|^2$ we can infer that $\nabla f(x_k)$ tends to zero, and we **morally** have a rate of convergence

$$\|\nabla f(x_k)\| = o\left(\frac{1}{\sqrt{k}}\right).$$

Also the above bounds is false in general, the optimizers have a trick to obtain such an estimate.

Theorem 6.3.3 Take any algorithm that have the two following properties :

- There exists $A \in \mathbb{R}$ such that $\sum_{k\geq 0}\|\nabla f(x_k)\|^2 \leq A$.
- The algorithms stops if $\|\nabla f(x_k)\| \leq \varepsilon$.

Then the algorithm stops before $\frac{A}{\varepsilon^2}$ iterations.

Proof

Suppose that the algorithm has spent n iterations, since the algorithm hasn't stopped, then $\|\nabla f(x_k)\| > \varepsilon$ for each $k \leq n$. It follows that

$$n\varepsilon^2 \leq \sum_{k=0}^{n-1}\|\nabla f(x_k)\|^2 \leq \sum_{k\geq 0}\|\nabla f(x_k)\|^2 \leq A.$$

And we have $n \leq \frac{A}{\varepsilon^2}$. The algorithm is then sure to stop before $A\varepsilon^{-2}$ iterations.

Theorem 6.3.4 As a application of Theorem 6.3.3, suppose that the hypothesis of Theorem 6.3.2 are true and consider a gradient algorithm with fixed step $s = \theta\frac{2}{L}$, with $0 < \theta < 1$. Suppose that the algorithms stops if $\|\nabla f(x_k)\| \leq \varepsilon$. Then the algorithm stops before $K\varepsilon^{-2}$ iterations. The exact value of K is given by

$$K = \frac{L}{2\theta(1-\theta)}\left(f(x_0) - \inf_{x\in\mathbb{R}^n} f(x)\right)$$

The huge problem with the above analysis is that we have no clear idea of the value of L.

6.4 Empiric Line search

6.4.1 A zoology of empiric line search

In this section, we suppose that the choice of direction of descent has been made and we focus on the choice of step.

This choice of step answers to two usually contradictory objectives, the first one is to find the best step possible (the one that decreases the most the function) and the second objective is to perform the smallest number of computations. At each end of the spectrum of compromise between the two different objectives, we find the **fixed step** and **optimal step** algorithms

FIXED STEP LINESEARCH.

```
$s_k = s_{k-1}$
```

OPTIMAL STEP LINESEARCH.

s_k is a solution of $\min_{s>0} f(x_k + sd_k)$

We will see below than none of these two strategies is very convincing. The first one is very risky, if the step is not chosen small enough, the algorithm may not converge. The second one is difficult to implement in practice. It amounts to solve a $1d$ optimization problem at each iteration. Moreover, the optimal step linesearch may be a total waste of computing power, why would we spend resources in finding a minimum in a direction that has no reason to be the correct one?

In this section, we describe some improvements of the fixed step and optimal step method. The first issue is the possible lack of convergence of the fixed step linesearch if the step is too large. We can enforce convergence by using the **backtracking** algorithm.

BACKTRACKING LINESEARCH.

```
$s_k = s_{k-1}$
while $f(x_k + s_k d_k) \geq f(x_k)$ :
    $s_k = s_k/2$
```

The backtracking linesearch doesn't allow the algorithm to augment the step if it was chosen to small. This algorithm can be modified such that it automatically augments the step before the backtracking :

BACKTRACKING WITH AUGMENTATION LINESEARCH.

```
$s_k = 1.3 * s_{k-1}$
while $f(x_k + s_k d_k) \geq f(x_k)$ :
    $s_k = s_k/2$
```

In the above algorithm, we take good care to augment the step with a factor different from the reduction, or else the algorithm could enter loops.

A plebiscited improvement is to try to perform an optimal linesearch but on a limited set of steps. We fiw in advance a set of possible multipliers to the step and we choose the best one.

PARTIAL LINESEARCH.

Input: s_{k-1}, $(a_i)_{1\leq i\leq m}$ an array of positive numbers with $a_1 = 1$.
Output: s_k.

- $S = \{a_i s_{k-1} \text{ for } 1 \leq i \leq m\}$.
- $s_k = \underset{s\in S}{\arg\min} f(x_k + s d_k)$.

It is important to chose numbers which are greater and smaller than 1 to let the algorithm decide between augmentation or diminution of the step.

6.4.2 The problems with empiric line search

Asserting that the function decreases is a desirable property but it may not be enough. We first discuss two examples where the iterates fail to converge. The first algorithm fails to converge because the steps are too big and the second one because the step is too small

Exercice 6.2: Too big/too small steps

Consider the function $f : x \mapsto \frac{1}{2}x^2$ with direction of descent given by $d_k = -x_k$. Assume that $x_0 = 2$.

1. Consider the choice of step
$$s_k = |x_k|^{-1}(2 + \frac{3}{2^{k+1}}).$$
Then $f(x_k)$ is decreasing but the algorithm does not converge in any sense (not to a critical point, not to a local minimum and not for the iterates). In this example, the step is **too big**.
2. Consider now the choice of step
$$s_k = \frac{1}{|x_k|2^{k+1}}.$$
Then the sequence $(f(x_k))_k$ is decreasing, the iterates converge but not to a critical point. In this example, the step is **too small**.

Solution to Exercice 6.2

1. In the first case, we have $x_{k+1} = x_k - \frac{x_k}{|x_k|}\left(2 + \frac{3}{2^{-k+1}}\right)$. Starting at $x_0 = 2$, a simple recurrence show that
$$x_k = (-1)^k(1 + \frac{1}{2^k}).$$
For every $k \in \mathbb{N}$: $f(x_{k+1}) < f(x_k)$. Hence we have a descent algorithm but the sequence $(x_k)_{k\in\mathbb{N}}$ does not converges, it has two accumulation points at $x = 1$ and $x = -1$. None of those two points correspond to extrema of f.
2. In the second case, use $x_{k+1} = x_k - \frac{x_k}{|x_k|}\left(\frac{1}{2^{k+1}}\right)$. Starting at $x_0 = 2$, a simple recurrence show that
$$x_k = 1 + \frac{1}{2^k}.$$
For every $k \in \mathbb{N}$: $f(x_{k+1}) < f(x_k)$. Hence we have a descent algorithm but the sequence $(x_k)_{k\in\mathbb{N}}$ converges to 1 which is not even a critical point of f.

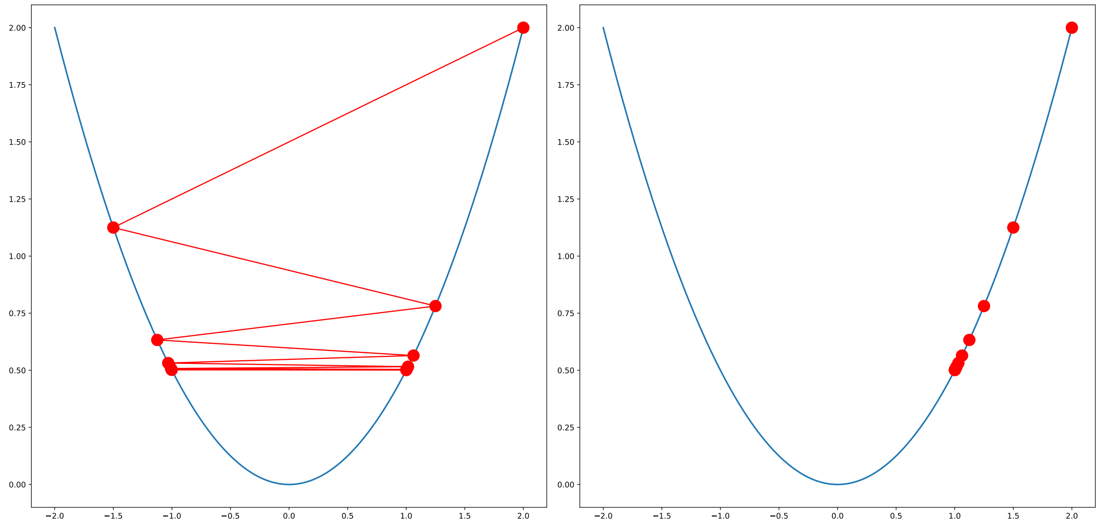


Figure 6.2: Examples of big and small steps

6.5 Wolfe line search

In order to circonvene the problems of small steps and of big steps, we impose two conditions on the steps, on which avoids small steps and one which avoids large steps. These conditions are called **Wolfe's** condition

Definition 6.5.1 — Wolfe's conditions. Let ε_1 and ε_2 be such that $0 < \varepsilon_1 < \varepsilon_2 < 1$, and d_k a direction of descent. We search a step s_k that verifies

- **Avoid big steps :**

$$f(x_k + s_k d_k) \leq f(x_k) + \varepsilon_1 s_k \langle \nabla f(x_k), d_k \rangle \tag{6.4}$$

- **Avoid small steps :**

$$\langle \nabla f(x_k + s_k d_k), d_k \rangle \geq \varepsilon_2 \langle \nabla f(x_k), d_k \rangle \tag{6.5}$$

The large step condition imposes that f decreases at least as much as ε_1 times its linear model. The small step condition imposes that ∇f is closer to 0, we get closer to a local minimum. It is easy to check that the second condition avoids small steps. Indeed the second condition is not valid for the choice $s_k = 0$. By continuity, it is not valid for too small a choice of s_k. In practice, we take $\varepsilon_1 = 10^{-4}$ and $\varepsilon_2 = 0.99$.

Proposition 6.5.1 — Notation with the merit function. In the proof, we introduce the following notation :

$$\phi : s \mapsto f(x_k + s d_k) - f(x_k).$$

We have $\phi(0) = 0$ and Wolfe's algorithm can be rewritten as :

$$\phi'(s) \geq \varepsilon_2 \phi'(0) \text{ and } \phi(s) \leq \varepsilon_1 s \phi'(0)$$

Proof

The only difficulty is to obtain the correct formula for $\phi'(s)$, we have

$$\phi'(s) = \langle \nabla f(x_k + sd_k), d_k \rangle$$

Indeed, performing a Taylor expansion, we have

$$\begin{aligned} \phi(s+h) &= f(x_k + (s+h)d_k) - f(x_k) \\ &= f(x_k + sd_k) + \langle \nabla f(x_k + sd_k), hd_k \rangle + o(h) - f(x_k) \\ &= \phi(s) + h \underbrace{\langle \nabla f(x_k + sd_k), d_k \rangle}_{=\phi'(s)} + o(h) \end{aligned}$$

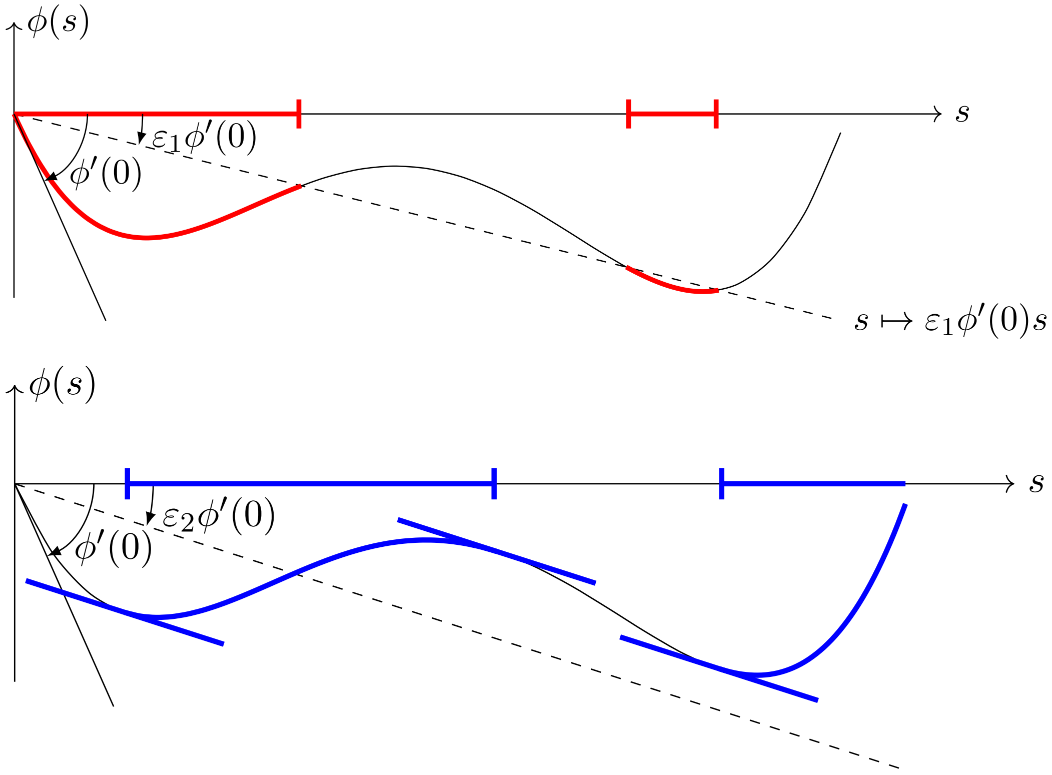


Figure 6.3: The Wolfe conditions. Top, the steps s such that $\phi(s) \leq \varepsilon_1 s\phi'(0)$. Bottom, the steps such that $\phi'(s) \geq \varepsilon_2\phi'(0)$.

6.5.1 Computation of a Wolfe step

The first thing to prove is that there exists a Wolfe step.

Proposition 6.5.2 — There exists a Wolfe step. Let f be differentiable and bounded from below, let d be a direction of descent of f at x, then there exists a step s that verifies Wolfe's conditions.

Proof

Let $\mathcal{A} = \{\eta > 0 \text{ s.t. } \forall 0 < r \leq \eta, \phi(r) \leq \varepsilon_1 r \phi'(0)\}$. Since $\varepsilon_1 < 1$, by continuity of $z(r) = \frac{\phi(r)}{r}$ with $z(0) = \phi'(0)$, then $\mathcal{A} \neq \emptyset$. Denote $s = \sup \mathcal{A}$. Since f is bounded from below, $s \neq +\infty$. For each n large enough such that $s - \frac{1}{n} > 0$, there exists $r_n \in [s - \frac{1}{n}, s]$ such that $\phi(r_n) \leq \varepsilon_1 r_n \phi'(0)$. We let n go to infinity and we have

$$\phi(s) \leq \varepsilon_1 s \phi'(0)$$

Hence s verifies the first Wolfe condition. For each n, $s + \frac{1}{n}$ does not belong to $\mathcal{A}$, so that there exists r_n with $s < r_n \leq s + \frac{1}{n}$ such that

$$\phi(r_n) > \varepsilon_1 r_n \phi'(0)$$

Let n goes to infinity, then $r_n \to s$ and $\phi(s) = \varepsilon_1 s \phi'(0)$. Then

$$\phi(r_n) - \phi(s) > \varepsilon_1 (r_n - s) \phi'(0)$$

Divide by $(r_n - s)$ which is positive, we obtain

$$\frac{\phi(r_n) - \phi(s)}{r_n - s} > \varepsilon_1 \phi'(0).$$

We let n go to infinity to obtain :

$$\phi'(s) \geq \varepsilon_1 \phi'(0) \geq \varepsilon_2 \phi'(0).$$

The simplest algorithm that computes a Wolfe step is attributed to Fletcher (1980) and Lemaréchal (1981), it is described here.

```
WOLFE LINESEARCH.
Input: $f$ a $C^1$ function, $x \in \mathbb{R}^n$ the actual point, $d$ the direction of descent of $f$ at $x$, $s_0$
       guess of Wolfe step, $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that : $0 < \varepsilon_1 < \varepsilon_2 < 1$
       .
Output: A step $s^\star$ that verify Wolfe conditions.
    1. $k := 0$; $s_- = 0$; $s_+ = +\infty$;
    2. While $s_k$ does not meet Wolfe condition
        (a) Too large : If $s_k$ does not meet (6.4):
                $s_+ = s_k \quad \text{and} \quad s_{k+1} = \frac{s_- + s_+}{2}$.
        (b) Too small If $s_k$ meets (6.4) but not (6.5) :
                $s_- = s_k \quad \text{and} \quad s_{k+1} = \begin{cases} \frac{s_- + s_+}{2} & \text{if } s_+ < +\infty \\ 2 s_k & \text{else.} \end{cases}$
        (c) $k := k + 1$;
    3. Return $s^\star = s_k$.
```

We admit without proof that under the exact same hypothesis of Proposition 6.5.2, the Wolfe linesearch ends in a finite number of iterations (provided no numerical error is made). The above algorithm is a very simple dichotomic interpolation and is rather slow. At each iteration, a linesearch is performed, hence it is of the essence to speed up the process. Remark that through the iterations k, the algorithm computes the values of $\phi(s_k)$ and $\phi'(s_k)$ (the merit function). The idea is to use these values to interpolate the merit function by a spline. In labwork, we will study the cubic spline interpolation method.

6.5.2 Convergence of descent method with Wolfe linesearch

We are interested in the convergence of any descent method with a Wolfe linesearch. We mainly show that

$$\lim_{k\to+\infty} \|\nabla f(x_k)\| = 0.$$

This results means that any accumulation point $\bar{x}$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a critical point of f (i.e. that $\nabla f(\bar{x}) = 0$).

Even if any accumulation point of the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a critical point, we do not say anything about the convergence of the sequence $(x_k)_k$. Indeed, there exists counterexamples, see [**Bertsekas99**] or [**NocedalWright**] for C^1 or C^2 functions. Recently, P.A. Absil, R. Mahoney et B. Andrews [**Absil2005**] proved the convergence of the iterates when the functions are analytic.

Theorem 6.5.3 — Convergence of Wolfe algorithm. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, with Lipschitz gradient and bounded from below. Let $(x_k)_k$ be a sequence of points given by an algorithm that ensures Wolfe condition are true

$$x_{k+1} = x_k + s_k d_k,$$

where $\langle d_k, \nabla f(x_k)\rangle < 0$. If $\cos(\theta_k) = \frac{\langle -\nabla f(x_k), d_k\rangle}{\|\nabla f(x_k)\|\|d_k\|}$, denotes the angle between d_k and $-\nabla f(x_k)$, then

$$\sum \cos(\theta_k)^2 \|\nabla f(x_k)\|^2 \text{ converges.}$$

Proof

The Wolfe condition : $\langle \nabla f(x_{k+1}), d_k\rangle \geq \varepsilon_2 \langle \nabla f(x_k), d_k\rangle$, yields

$$\langle \nabla f(x_{k+1}) - \nabla f(x_k), d_k\rangle \geq (\varepsilon_2 - 1)\langle \nabla f(x_k), d_k\rangle.$$

Moreover, we have

$$\begin{array}{rcl}\langle \nabla f(x_{k+1}) - \nabla f(x_k), d_k\rangle & \leq & \|\nabla f(x_{k+1}) - \nabla f(x_k)\|\|d_k\| \\ & \leq & L\|x_{k+1} - x_k\|\|d_k\| = L s_k \|d_k\|^2.\end{array}$$

Combining the two inequalities, we have

$$s_k \geq \frac{\varepsilon_2 - 1}{L} \frac{\langle \nabla f(x_k), d_k\rangle}{\|d_k\|^2} > 0.$$

Using the first Wolfe condition, we then have

$$\begin{array}{rcl} f(x_k) - f(x_{k+1}) & \geq & -\varepsilon_1 s_k \langle \nabla f(x_k), d_k\rangle \geq \varepsilon_1 \frac{1-\varepsilon_2}{L} \frac{\langle \nabla f(x_k), d_k\rangle^2}{\|d_k\|^2} \\ & \geq & \varepsilon_1 \frac{1-\varepsilon_2}{L} \cos(\theta_k)^2 \|\nabla f(x_k)\|^2 \geq C \cos(\theta_k)^2 \|\nabla f(x_k)\|^2 \end{array}$$

Summing up these inequalities, we obtain for each n,

$$f(x_0) - \inf_{x\in\mathbb{R}^n} f(x) \geq f(x_0) - f(x_{n+1}) \geq C \sum_{k=0}^{n} \cos(\theta_k)^2 \|\nabla f(x_k)\|^2.$$

$$f(x_0) - \inf_{x\in\mathbb{R}^n} f(x) \geq C \sum_{k=0}^{+\infty} \cos(\theta_k)^2 \|\nabla f(x_k)\|^2.$$

Note that θ_k is the the angle between the direction of descent d_k and the direction $-\nabla f(x_k)$, meaning that the choice $d_k = -\nabla f(x_k)$ is the one that optimizes the convergence rate. This is consistent with the idea that the gradient is the steepest descent direction. Note also that if we ensure that the cosinus is bounded from below by a strictly positive constant (i.e. the direction d_k does not become too much orthogonal to $-\nabla f(x_k)$, then there exists a constant A such that

$$\sum_{k\geq 0} \|\nabla f(x_k)\|^2 \leq A.$$

We can deduce the following theorem :

Theorem 6.5.4 As a corollary of Theorem 6.3.3, consider a descent algorithm with Wolfe linesearch such that

- There exists a $c > 0$ with $\cos(\theta_k) > c$ for all k
- The algorithms stops if $\|\nabla f(x_k)\| \leq \varepsilon$.

Then the algorithm stops before $K\varepsilon^{-2}$ iterations. The exact value of K is given by

$$K = \frac{L}{c^2\varepsilon_1(1-\varepsilon_2)} \left(f(x_0) - \inf_{x\in\mathbb{R}^n} f(x) \right)$$

Proof

Noting that $\|\nabla f(x)\|^2 \leq K$, the proof is an application of Theorem 6.3.3 with $A = K$.

6.5.3 Comparison of first order methods

In this section, we study the performance of the linesearch methods for the following function

$$f : M = (x, y) \in \mathbb{R}^2 \mapsto \frac{1}{2}x^2 + \frac{7}{2}y^2,$$

The function f is a C^2 function whose minimum is attained at $M^\star = (0, 0)$ (only critical point). The function f is a strictly convex function. We denote $M_k = (x_k, y_k)$ the current iterate. The descent direction is then given by

$$d_k = -\nabla f(M_k) = \begin{pmatrix} -x_k \\ -7y_k \end{pmatrix}.$$

- **Fixed step strategy :** One can check easily that ∇f is L-Lipschitz with $L = 5\sqrt{2}$. This gives an upper bound on the step $\frac{2}{L} \simeq 0.2828$ In table **??**, we give the result of the algorithm for different values of the step.
- **Optimal step strategy :** At each iteration the optimal step strategy amounts to solve

$$\min_{s>0} f(M_k + sd_k) = \frac{1}{2}x_k^2(1-s)^2 + \frac{7}{2}y_k^2(1-7s)^2$$

step	0.325	0.25	0.125	0.05	0.01
iteration number	DV	49	101	263	1340

Table 6.1: Number of iterations of the fixed step gradient algorithm in order to approach a critical point of f within 10^{-5} accuracy. Initial point $x_0 = (7, 1.5)$.

The above function is a second order polynom in s, the solution of the above problem is given by:

$$s_k = \frac{x_k^2 + 7^2 y_k^2}{x_k^2 + 7^3 y_k^2}.$$

In order to reach a critical point with 10^{-5} accuracy, starting with $x_0 = (7, 1.5)$, the algorithm requires 43 iterations.

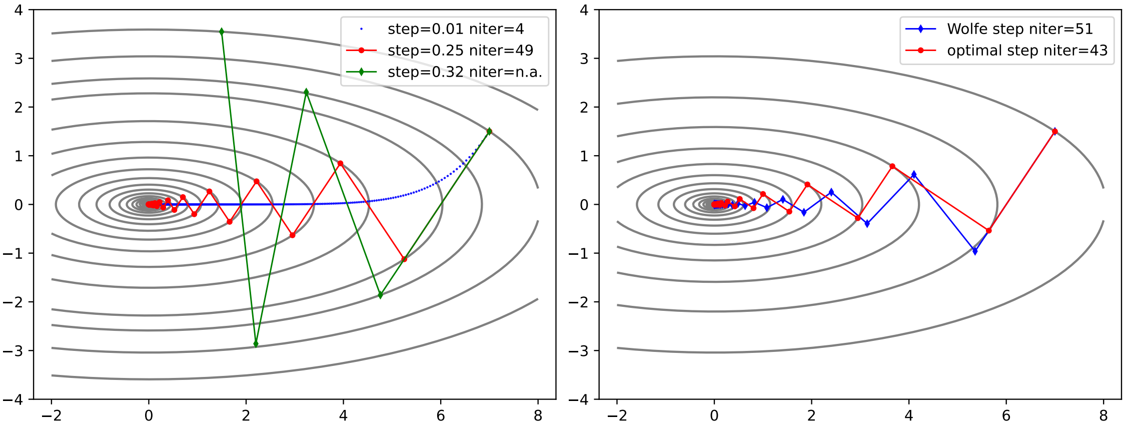


Figure 6.4: Gradient algorithm for a quadratic function with initial point $x_0 = (7, 1.5)$. On the Left, fixed step algorithm and on the right : Steepest descent algorithm (red) and Wolfe linesearch algorithm (blue).

On Figure 6.5.3, we display the caracteristic behavior of the methods with fixed step or optimal step, they are:

- Optimal-step algorithm is slow to converge, because the directions are orthogonal to each other.
- Fixed-step algorithms might not converge.

CHAPTER 7

Second order descent methods

7.1 Newton algorithms

7.1.1 Choice of descent direction and of step: Newton algorithm

The Newton algorithm is the most basic algorithm of second order. It exhibits very high rates of convergence... when it converges. It is a very fast algorithm which is not very stable and which requires heavy computations per iterations. We first state the Newton algorithm and then we give three interpretations of this algorithm.

> **Definition 7.1.1 — Newton's algorithm.** The Newton algorithm for the unconstrained minimization of $f : \mathbb{R}^n \mapsto \mathbb{R}$ reads as follows. Start with x_0 and for each iteration k, do
>
> 1. $d_k = -H[f](x_k)^{-1}(\nabla f(x_k))$
> 2. $x_{k+1} = x_k + d_k$

The Newton algorithm is an algorithm with step $s_k = 1$ and a direction d_k which is almost the one of the gradient algorithm. Indeed the direction for the gradient algorithm is given by $-\nabla f(x_k)$ and in order to obtain the one for the Newton 's algorithm, it is sufficient to multiply it by the inverse of the Hessian. We can understand immediatly one of the main limitations of the Newton's algorithm. Suppose that instead of minimizing the function f, a student aims at **maximizing** the function f. A good idea is then to minimize the function $-f$. If the student writes down the corresponding algorithm and denote $\tilde{d}_k$ its update direction, he finds that

$$\tilde{d}_k = -H[-f](x_k)^{-1}(\nabla(-f)(x_k)) = -H[f](x_k)^{-1}(\nabla f(x_k)) = d_k.$$

Hence, the Newton algorithm for the minimization of a function or for the maximization of the same function is the same !!! Hence there is no way to tell if the Newton algorithm is used as a minimization or a maximization algorithm !! In order to ensure that the algorithm is indeed a minimization algorithm, a good hypothesis is to check that the direction d_k is a direction of descent. This is ensured by the following proposition

> **Proposition 7.1.1** Suppose that $H[f](x_k) > 0$, then $H[f](x_k)^{-1}$ exists and the Newton's algorithm is doable and the direction d_k given by $d_k = -H[f](x_k)^{-1}(\nabla f(x_k))$is

a direction of descent, provided that $\nabla f(x_k) \neq 0$.

Proof

If $H[f](x_k) > 0$ then this matrix has no zero eigenvalue and hence is invertible. If $(\lambda_i)_i$ denote the eigenvalues of $H[f](x_k)$, the eigenvalues of $H[f](x_k)^{-1}$ are then given by $(\frac{1}{\lambda_i})_i$. Hence $H[f](x_k)^{-1} > 0$. Denote $A = H[f](x_k)^{-1}$, we have, if $\nabla f(x_k) \neq 0$

$$\langle d_k, \nabla f(x_k)\rangle = \langle -A\nabla f(x_k), \nabla f(x_k)\rangle < 0 \text{ because } A > 0.$$

Hence d_k is a direction of descent.

This algorithm has three different interpretations. We will see that each interpretation requires that $H[f](x_k) > 0$ in order to be able to conclude.

7.1.2 Newton Algorithm as a search for a critical point

The first interpretation of the Newton algorithm stems from a numerical algorithm used when solving non-linear equations. We recall this algorithm which is incidently also called Newton's algorithm.

Definition 7.1.2 — Newton algorithm for non-linear equations. Let $F : \mathbb{R}^n \mapsto \mathbb{R}^n$. The following algorithm aims at solving the non-linear system of n equations with n unknowns given by $F(x) = 0$

- $d_k = -(Jac_{x_k}[F])^{-1}F(x_k)$
- $x_{k+1} = x_k + d_k$

The idea of this algorithm is as follows : suppose that we are close to the solution and that we think there exists a small h such that $F(x_k + h) = 0$. Then we perform a first order Taylor expansion and we find

$$F(x_k) + (Jac_{x_k}[F])h \simeq 0.$$

Or equivalently $h \simeq d_k$ if $d_k = -(Jac_{x_k}[F])^{-1}F(x_k)$. Since the critical point is at $x_k + h$, it makes sense to set $x_{k+1} = x_k + d_k$. We see in with this interpretation that Newton's algorithm is not an algorithm that aims at finding a minimizer but a critical point. That's why Newton's algorithm can also be interpreted as a maximization algorithm since maximizer are also critical point. The question is to ensure that the critical point is a minimizer. Using Euler's condition, if x_k is a critical point, it is sufficient to suppose that $H[f](x_k) > 0$ in order to ensure that x_k is a local minimizer.

7.1.3 Newton Algorithm as a second order expansion

Suppose that $H[f](x_k) > 0$. The problem can be rephrased into finding h a solution of

$$\min_d f(x_k + d),$$

and to set $x_{k+1} = x_k + d$. Just as the previous section, assume that d is small and replace $f(x_k + d)$ by its second order Taylor expansion. Then the problem becomes

$$\min_d f(x_k) + \langle \nabla f(x_k), d\rangle + \frac{1}{2}\langle H_{x_k}[f]d, d\rangle$$

This problem is a quadratic problem, provided that the matrix $H[f](x_k)$ is positive definite, it admits a unique solution denoted d_k given by

$$H_{x_k}[f]d_k = -\nabla f(x_k)$$

Then one obtains

- $d_k = -H[f](x_k)^{-1}(\nabla f(x_k))$
- $x_{k+1} = x_k + d_k$

Intuitively, we undertand that Newton's algorithm should be better than the gradient method, because the gradient method is based on a first order Taylor expansion and aims at solving

$$\min_{\|d\|\leq m} f(x_k) + \langle \nabla f(x_k), d\rangle$$

Note also that quadratic problem can be minimized if and only if the matrix is positive definite. We see here that $H[f] \succ 0$ is a condition for the Newton algorithm to be efficient.

7.1.4 Newton Algorithm as a trust algorithm

The Hessian $H[f](x_k)$ is a symetric matrix and hence there exists an orthonormal basis of eigenvectors of $H[J](v_k)$. Denote $(e_i)_i$ this basis and $(\lambda_i)_i$ the corresponding eigenvalues. Then λ_i represents :

- The rate of change of $\nabla f(x_k)$ in direction e_i.
- The higher $|\lambda_i|$, the less the value of $\nabla f(x_k)$ in direction e_i can be trusted.

Recall that the gradient method amounts to take, as a direction of descent

$$d_k = -\nabla f(x_k) \sum_i \langle -\nabla f(x_k), e_i\rangle e_i$$

And the Newton method amounts to take

$$d_k = H[f](x_k)^{-1}(-\nabla f(x_k) = \sum_i \frac{1}{\lambda_i}\langle -\nabla f(x_k), e_i\rangle e_i.$$

So that, in order to retrieve Newton's method, one has to follow a gradient method and to divide each component of the direction of descent by λ_i. When $H[f] \succ 0$, each λ_i is > 0, so that Newton's method amounts to divide the components by a factor that represents the **mistrust** in the corresponding direction. Put another way, in the Newton's method, the more you can **trust** a direction e_i, the further you go in this corresponding direction. Of course this interpretration breaks totally if an eigenvalue λ_i is < 0, in this case, the Newton method goes in the opposite direction for e_i !!!

7.1.5 Newton : Pros and cons

We apply Newton's method to the following problems

$$\begin{array}{ll} (P_1) & \min_{(x,y)\in\mathbb{R}^2} f(x,y) = 100(y-x^2)^2 + (1-x)^2 \text{ (Rosenbrock)}. \\ (P_2) & \min_{(x,y)\in\mathbb{R}^2} g(x,y) = \frac{1}{2}x^2 + x\cos ymbox(Oscill). \end{array}$$

The problem (P_1) admits a unique critical point at $(1,1)$ which is a global minimum of f, whereas problem (P_2) admits an infinite number of critical points :

$$\begin{array}{ll} ((-1)^{k+1}, k\pi),\ k\in\mathbb{Z} & \text{local minima of } g \\ (0, \frac{\pi}{2}+k\pi),\ k\in\mathbb{Z} & \text{saddle-points of } g \end{array}$$

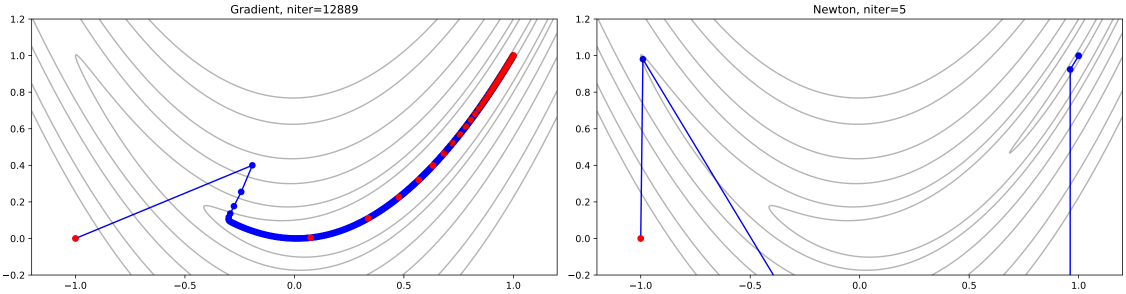


Figure 7.1: Rosenbrock:

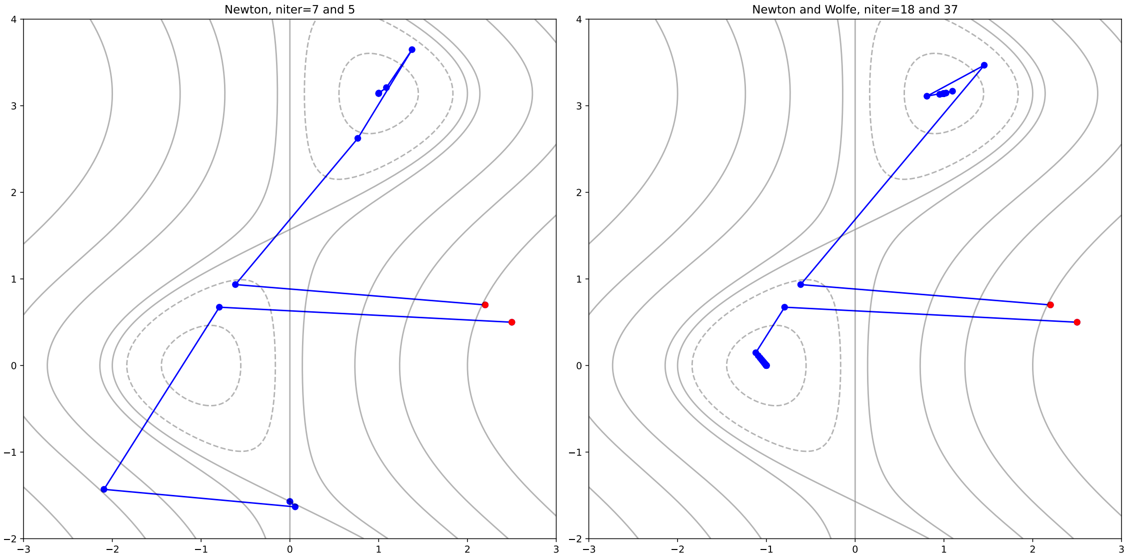


Figure 7.2: Oscill:

In Figure 7.3, we show the results for the optimization of the Rosenbrock function and in Figure 7.4 for the Oscill function.

These figures showcases the assets and liabilities of convergence of the Newton method

- Pros
  1. This algorithm converges quadratically (multiply by 2 the number of decimal of precision at each iteration)
  2. Quadratic problems converge in one iteration.
  3. The step is 1.
  4. If $H[f](x_k) > 0$, it is an algorithm of descent.
- Cons
  1. Compute the Hessian $H[f](x_k)$
  2. Solve at each iteration $H[f](x_k)^{-1}(-\nabla f(x_k))$.
  3. Must start close to the minimum (basin of attraction).
  4. Not possible to compute d_k if the Hessian is singular.
  5. No distinction made between minima, maxima and saddle points.

7.1.6 Convergence of Newton

The Newton's method enjoys quadratic convergence, the result is made precise in the following theorem.

Theorem 7.1.2 Let x_k follow a Newton algorithm. Assume $f : \mathbb{R}^n \mapsto \mathbb{R}$ is C^2 with M-Lipschistz Hessian and suppose there exists γ such that for all x, $H[f](x) \succeq \gamma Id$. Denote $a_k = \frac{M}{2\gamma^2}\|\nabla f(x_k)\|$, then

$$a_{k+1} \leq a_k^2.$$

Especially, if x_0 is close enough to a critical point of f so that $a_0 < 1$, then $\|\nabla f(x_k)\|$ goes quadratically fast towards 0.

Proof

We have

$$\|\nabla f(x_{k+1}) - \nabla f(x_k) - Hf[x_k](x_{k+1} - x_k)\| \leq \frac{M}{2}\|x_{k+1} - x_k\|^2$$

Replace $x_{k+1} - x_k$ by $H[f](x_k)^{-1}\nabla f(x_k)$ to obtain

$$\|\nabla f(x_{k+1})\| \leq \frac{M}{2}\|H[f](x_k)^{-1}\nabla f(x_k)\|^2 \leq \frac{M}{2\gamma^2}\|\nabla f(x_k)\|^2$$

In a nutshell, if the Newton method is already closed to a critical point and if $H[f](x) \succeq \gamma Id$, then we obtain quadratic convergence. But if we start too far away from a critical point, we might not have convergence. A good example is given by the following exercise

Exercice 7.1

Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \sqrt{1 + x^2}$. Show that f is strongly convex, that it verifies the hypothesis of Theorem 7.1.2 but that it fails to converge whenever $|x_0| > 1$.

However, the Wolfe linesearch helps stabilizing the Newton method, thanks to the following proposition

Proposition 7.1.3 Let A be a symmetric definite positive matrix with $\lambda_0 \leq \dots \lambda_n$. Denote κ the 2-conditionning number of A, we recall that it is defined as:

$$\kappa = \|A\|_{2\to 2}\|A^{-1}\|_{2\to 2} = \frac{\lambda_n}{\lambda_0}$$

then for every vector $u \neq 0$, we have

$$\frac{\langle Au, u\rangle}{\|u\|\|Au\|} > \frac{1}{\sqrt{\kappa}}$$

Proof

Let $(e_i)_i$ be an orthonormalised basis of eigenvectors of A and denote u_i the coordinates of u, we have $u = \sum u_i e_i$ and $Au = \sum \lambda_i u_i e_i$. It follows that

$$\langle Au, u\rangle = \sum_i \lambda_i u_i^2$$

We have

$$\begin{aligned}
\|Au\|^2 &= \sum_i \lambda_i^2 u_i^2 \leq \lambda_n \sum_i \lambda_i u_i^2 = \lambda_n \langle Au, u\rangle \\
\|u\|^2 &= \sum_i u_i^2 = \sum_i \frac{\lambda_i}{\lambda_i} u_i^2 \\
&\leq \frac{1}{\lambda_0}\sum_i \lambda_i u_i^2 = \frac{\langle Au, u\rangle}{\lambda_0}
\end{aligned}$$

Finally

$$\frac{\langle Au, u\rangle^2}{\|u\|^2\|Au\|^2} \geq \frac{\lambda_0}{\lambda_n} = \frac{1}{\kappa}$$

Proposition 7.1.4 Use a Wolfe linesearch and suppose that the 2-conditionning number of the Hessian of f is uniformely bounded through the iterations, that is

$$\exists M > 0 \text{ such that } \forall k, \|H[f](x_k)\|_{2\to 2}\|H[f](x_k)^{-1}\|_{2\to 2} \leq M.$$

Suppose that the Hessian of f is positive through the iterations, then the Newton algorithm with Wolfe step converges.

In Figure 7.3, we show the results for the optimization of the Rosenbrock function and in Figure 7.4 for the Oscill function.

7.2 Quasi-Newton algorithm

7.2.1 Defintion

The class of **Quasi-Newton** algorithms is a class of algorithms such that the direction of descent d_k is computed by

$$d_k = H_k^{-1}(-\nabla f(x_k)),$$

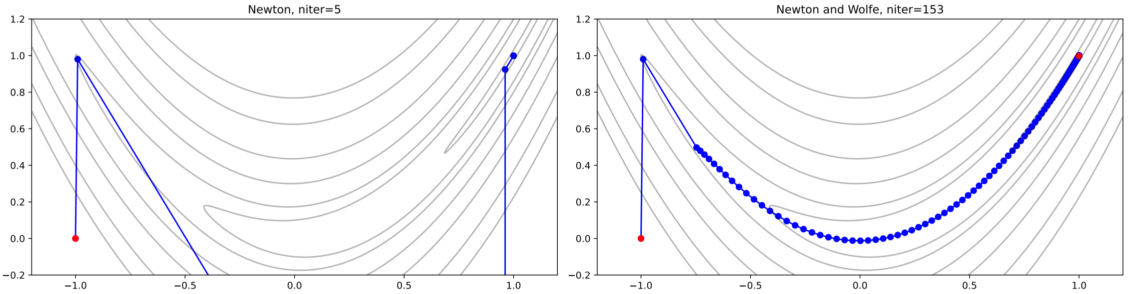


Figure 7.3: Rosenbrock:

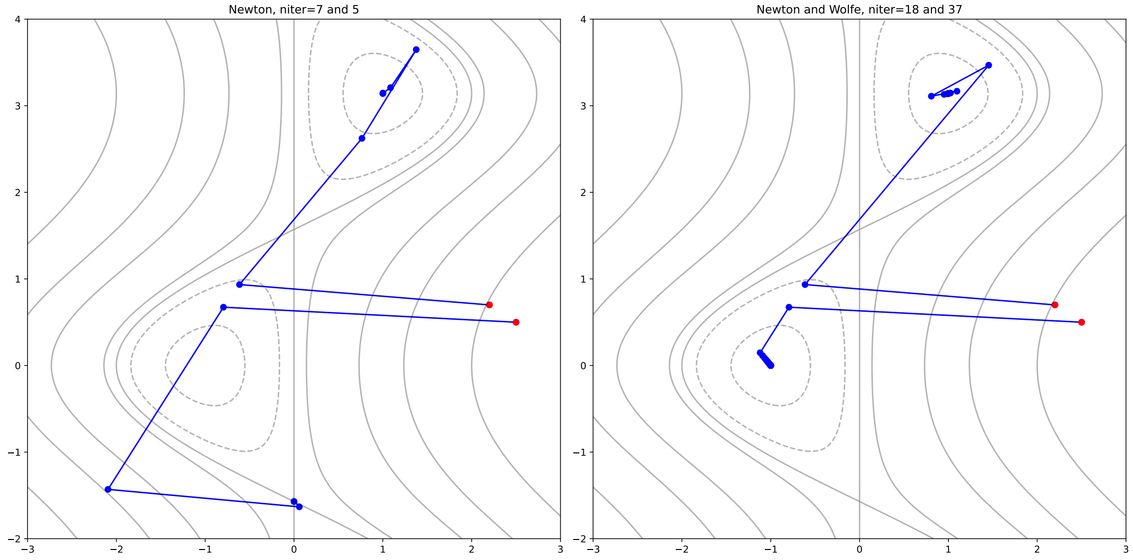


Figure 7.4: Oscill:

where H_k is a positive definite matrix. The most basic Quasi-Newton algorithm is... the gradient algorithm... which amounts to set $H_k = Id$. This is the most famous optimizer joke (and surely the best one... sadly :( ), in practice, the optimizer will try to design a matrix H_k which is the closest possible to the Hessian while being positive definite. The first naïve idea is to set

$$H_k = H[f](x_k) + \alpha Id,$$

with $\alpha \geq 0$. Then H_k is equal positive definite if and only if α is greater than $-\lambda_0$, the smallest eignvalue of $H[f](x_k)$. Setting α cloes to 0, yields an Newton algorithm and setting α very large yields $H_k \simeq \alpha Id$, and hence we retrieve a gradient algorithm. Intermediate values of α can be seen as a mixture between Newton's algorithm and the gradient algorithm. In practice, if $\lambda_0 > 0$, setting $\alpha = 0$ is preferable.

7.2.2 Gauss-Newton algorithm

A very important class of problems is the "least-square" problem that appears in data mining, inverse problems, statistical analysis, learning. It is a problem that can be stated the following way :

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2} \sum_{i=1}^{m} F_i(x)^2, \tag{7.1}$$

where F is a map from $\mathbb{R}^n$ to $\mathbb{R}^m$, with $m \geq n$.

This problem arises when one wants to find solution of $F(x) = 0$ and when this possibly non-linear system is overdetermined.

The gradient and Hessian of f is given from the following proposition

Proposition 7.2.1 Let $F : \mathbb{R}^n \mapsto \mathbb{R}^p$, and let $f : \mathbb{R}^n \to R$ be equal to $f(x) = \frac{1}{2}\|F(x)\|^2$, then

$$\begin{array}{rcl} \nabla f(x) & = & (Jac[F](x))^T F(x) = \sum_{i=1}^m F_i(x) \nabla F_i(x) \\ H[f](x) & = & (Jac[F](x))^T Jac[F](x) + \sum_{i=1}^m F_i(x) H[F_i](x) \end{array}$$

The Gauss-Newton algorithm takes a step $s_k = 1$ in the direction of descent :

$$d_k = H_k^{-1}(-\nabla f(x_k)) \quad H_k = (Jac[F](x_k))^T Jac[F] x_k$$

The approximation of the Hessian H_k is constructed from the Hessian by dropping the term $a_k = \sum_{i=1}^m F_i(x_k) H[F_i](x_k)$. There are several reasons to perform this operation

- **Laziness:** This term a_k is quite difficult to compute, indeed it as second order derivatives of F inside. The term which is kept only has first order derivatives.
- **The term is small anyway:** Suppose that x_k is close to the minimizer and suppose that $F(x_k) \simeq 0$, then the term a_k is close to zero also because each $F_i(x_k)$ is close to zero. Hence, it is ok to discard this term
- **It is convenient:** By construction H_k is symetric and $H_k \succeq 0$. Indeed if A is any matrix, then $A^T A$ is symetric and $A^T A \succeq 0$, because for every vector h, we have

$$\langle A^T A h, h \rangle = \langle Ah, Ah \rangle = \|Ah\|^2$$

7.2.3 An other interpretation of Gauss-Newton's algorithm

Take a least-square problem and suppose that F is linear, that is there exists a matrix A and a vector b such that $F(x) = Ax - b$, then we obtain the so-called linear least-square problem:

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

We recall that the linear-least square problem admits solution which are given by the so-called **normal** equations

Proposition 7.2.2 Let $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, $b \in \mathbb{R}^m$ and $m \leq n$. Suppose that A is of **full-rank**, that is $rank(A) = n$. Then the least-square problem

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

admits a unique minimizer $x^\star$ given by $x^\star = \left(A^TA\right)^{-1} A^Tb$

7.3 The BFGS Algorithm

The BFGS algorithm aims at computing an approximation of the Hessian without extra computations. We suppose that for a sequence of points $(x_k)_k$, we have access to $\nabla f(x_k)$. Because the Hessian of f is the Jacobian of the gradient, we must have, when x_{k+1} is close to x_k :

$$\nabla f(x_k) \simeq \nabla f(x_{k+1}) + H[f](x_{k+1})(x_k - x_{k+1}).$$

We focus on algorithm for which the above approximation is an equality, we will then ensure that $H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$.

Proposition 7.3.1 The BFGS algorithm is an algorithm that will obey the following properties : It will construct at each iteration a matrix H_k which verifies

- H_k is symmetric, positive definite and at each iteration.
- If $\sigma_k = x_k - x_{k-1}$ and $y_k = \nabla f(x_k) - \nabla f(x_{k-1})$, then $H_k\sigma_k = y_k$.
- The choice of direction of descent is then $d_k = -H_k^{-1}\nabla f(x_k)$.

A necessary condition for the BFGS algorithm to be built is $(\sigma_k, y_k) > 0$.

Proof

If there exists a positive definite matrix such that $H_k\sigma_k = y_k$, then we must have $(H_k\sigma_k, \sigma_k) > 0$. Hence a necessary condition for the existence of a BFGS algorithm is $(\sigma_k, y_k) > 0$.

The rules of Proposition 7.3.1 do not define the matrix $H_k \in \mathcal{M}_{nn}(\mathbb{R})$, there are many matrices that verify Proposition 7.3.1. In order to add some extra rules, we suppose two things : First H_k is not too different from H_{k-1}, second, in order to enforce $H_k \succ 0$ and H_k symetric we suppose that there exists a square matrix S_k such that $H_k = S_kS_k^T$. For any vector g, we define $\tilde{S}$ the unique minimizer of

$$\min_{Sg=y_k} \frac{1}{2}\|S - S_{k-1}\|_2^2. \tag{7.2}$$

and we will show that it is possible to find g such that $g = \tilde{S}^T\sigma_k$ and we will set $S_k = \tilde{S}$. We now proceed to the construction:

Denote $\tilde{S}$ the unique minimum of (7.2). There exists a unique minimizer because Problem (7.2) is an orthogonal projection on an hyperplane. We rewrite the equality constraints as $-\langle S, e_ig^T\rangle = -\langle y_k, e_i\rangle$ for each $i = 1, \ldots, n$ and introduce the Lagrange multiplier $\lambda \in \mathbb{R}^n$ to obtain the KKT equations

$$\tilde{S} = S_{k-1} + \lambda g^T \tag{7.3}$$

$$\tilde{S}g = y \Longleftrightarrow \lambda = \frac{1}{\|g\|^2}(y_k - S_{k-1}g) \tag{7.4}$$

From (7.3), we deduce $\tilde{S}^T = S_{k-1}^T + g\lambda^T$ and if we want to verify $g = \tilde{S}^T\sigma_k$ then we must have

$$g = S_{k-1}^T\sigma_k + g\langle\lambda, \sigma_k\rangle. \tag{7.5}$$

Hence there must exists $\alpha \in \mathbb{R}$ such that $g = \alpha S_{k-1}^T\sigma_k$. Plugging $g = \alpha S_{k-1}^T\sigma_k$ into (7.3) and using (7.4) to remove λ, we obtain :

$$1 = \alpha(1 - \langle\lambda, \sigma_k\rangle) = \alpha\left(1 - \frac{\langle y_k, \sigma_k\rangle - \alpha\langle S_{k-1}S_{k-1}^T\sigma_k, \sigma_k\rangle}{\alpha^2\|S_{k-1}^T\sigma_k\|^2}\right),$$

the above equation is satisfied for $\alpha = \pm\frac{\sqrt{\langle y_k,\sigma_k\rangle}}{\|S_{k-1}^T\sigma_k\|}$. From (7.3), we now have the following formula for $S_k = \tilde{S}$

$$\begin{aligned} S_k &= S_{k-1} + \frac{(y_k - S_{k-1}g)g^T}{\|g\|^2} = S_{k-1} + \frac{(y_k - \alpha S_{k-1}S_{k-1}^T\sigma_k)\sigma_k^T S_{k-1}}{\alpha\|S_{k-1}^T\sigma_k\|^2} \\ &= S_{k-1} + \frac{(y_k - \alpha H_{k-1}\sigma_k)\sigma_k^T S_{k-1}}{\alpha\langle H_{k-1}\sigma_k, \sigma_k\rangle} \end{aligned}$$

We now compute $H_k = S_k S_k^T$ with

$$\begin{aligned} H_k &= S_k S_k^T \\ &= S_{k-1}S_{k-1}^T + \frac{(y_k - \alpha H_{k-1}\sigma_k)\sigma_k^T H_{k-1}}{\alpha\langle H_{k-1}\sigma_k, \sigma_k\rangle} + \frac{H_{k-1}\sigma_k(y_k - \alpha H_{k-1}\sigma_k)^T}{\alpha\langle H_{k-1}\sigma_k, \sigma_k\rangle} \\ &\quad + \frac{(y_k - \alpha H_{k-1}\sigma_k)(y_k - \alpha H_{k-1}\sigma_k)^T}{\alpha^2\langle H_{k-1}\sigma_k, \sigma_k\rangle} \\ &= H_{k-1} + \frac{y_k y_k^T}{\alpha^2\langle\sigma_k, H_{k-1}\sigma_k\rangle} - \frac{H_{k-1}\sigma_k\sigma_k^T H_{k-1}}{\langle\sigma_k, H_{k-1}\sigma_k\rangle}. \end{aligned}$$

This yields the following BFGS update:

Definition 7.3.1 — BFGS (Broyden, Fletcher, Goldfarb, Shannon. 1969-70). Suppose H_0 is a symetric definite positive matrix. Suppose that for each k, $\langle\sigma_k, y_k\rangle > 0$, the BFGS update is :

$$H_k = H_{k-1} + \frac{y_k y_k^T}{\langle y_k, \sigma_k\rangle} - \frac{H_{k-1}\sigma_k(H_{k-1}\sigma_k)^T}{\langle\sigma_k, H_{k-1}\sigma_k\rangle}.$$

Then for every k, H_k is symetric positive definite and its inverse $B_k = H_k^{-1}$ is defined by :

$$B_k = \left(I - \frac{\sigma_k y_k^T}{\langle y_k, \sigma_k\rangle}\right) B_{k-1}\left(I - \frac{y_k\sigma_k^T}{\langle y_k, \sigma_k\rangle}\right) + \frac{\sigma_k\sigma_k^T}{\langle y_k, \sigma_k\rangle}.$$

Moreover, we always have $H_k\sigma_k = y_k$ and $B_k y_k = \sigma_k$.

Proof

We have several things to prove, suppose that at iteration k, H_{k-1} is a symetric definite positive matrix with inverse given by B_{k-1}. We have to prove that the property holds for H_k and B_k.

- By a direct computation, $H_k\sigma_k = y_k$ and $B_k y_k = \sigma_k$.
- It is a direct computation to prove that $H_k^T = H_k$ and $B_k^T = B_k$.
- We prove now that B_k is definite positive. Denote C as $C = I - \frac{y_k\sigma_k^T}{\langle y_k,\sigma_k\rangle}$ then for any $x \neq 0$ and $u = Cx$,

$$\langle B_k x, x\rangle = \langle C^T B_{k-1} C x, x\rangle + \frac{\langle \sigma_k\sigma_k^T x, x\rangle}{\langle y_k,\sigma_k\rangle} = \langle B_{k-1}u, u\rangle + \frac{\langle \sigma_k, x\rangle^2}{\langle y_k,\sigma_k\rangle},$$

Since $\langle \sigma_k, y_k\rangle > 0$, we have $\langle B_k x, x\rangle \geq 0$ for all $x \in \mathbb{R}^n$. It remains to show that $\langle B_k x, x\rangle \neq 0$. Since $B_{k-1} \succ 0$, we have

$$\begin{aligned} \langle B_k x, x\rangle = 0 &\Rightarrow \langle B_{k-1}u, u\rangle \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \\ &\Rightarrow u = 0 \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \end{aligned}$$

but $u = Cx = x - \frac{y_k\langle x,\sigma_k\rangle}{\langle y_k,\sigma_k\rangle}$. Hence

$$u = 0 \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \Rightarrow x = 0$$

- We now prove that $H_k B_k = Id$. For any $e \in \mathbb{R}^n$, we decompose it as the sum of an element of $Vect(y_k)$ and $Vect(\sigma_k)^{\perp}$ using the formula

$$e = \frac{y_k\sigma_k^T}{\langle \sigma_k, y_k\rangle} e + (Id - \frac{y_k\sigma_k^T}{\langle \sigma_k, y_k\rangle})e.$$

Since we already know that $H_k B_k y_k = y_k$, it is sufficient to check that $H_k B_k u = u$ for all $u \in Vect(\sigma_k)^{\perp}$. Take such an u an remark that

$$B_k u = (Id - \frac{\sigma_k y_k^T}{\langle \sigma_k, y_k\rangle}) B_{k-1} u \tag{7.6}$$

so that $B_k u$ is orthogonal to y_k. Denoting $z = B_k u$, we obtain the following formula for $H_k z$:

$$H_k z = H_{k-1} z - H_{k-1}\sigma_k \frac{\langle H_{k-1}\sigma_k, z\rangle}{\langle H_{k-1}\sigma_k, \sigma_k\rangle} \tag{7.7}$$

And denoting $\alpha = \frac{\langle y_k, B_{k-1}u\rangle}{\langle \sigma_k, y_k\rangle}$,we have from(7.6) $z = B_k u = B_{k-1}u - \alpha\sigma_k$. We plug back this equation in (7.7) and we obtain

$$H_k z = H_{k-1} B_{k-1} u - H_{k-1}\sigma_k \frac{\langle H_{k-1}\sigma_k, B_{k-1}u\rangle}{\langle H_{k-1}\sigma_k, \sigma_k\rangle}$$

By recurence $H_{k-1}B_{k-1} = Id$ so that $\langle H_{k-1}\sigma_k, B_{k-1}u\rangle = \langle \sigma_k, u\rangle = O$, because $u \in Vect(\sigma_k)^{\perp}$. Finally we have

$$H_k B_k u = u.$$

Which proves that H_k and B_k are mutually inverse.

The BFGS algorithm yields an approximation of the Hessian H_k and a computation of its inverse B_k for each k. The only hypothesis needed for the BFGS algorithm to work is that algorithm is fed with vectors y_k and σ_k such that, for each k, we have : $\langle \sigma_k, y_k \rangle > 0$. Luckily, it turns out that BFGS algorithm works in pair with Wolfe's linesearch method. Indeed, BFGS algorithm requires $\langle \sigma_k, y_k \rangle > 0$ for B_k to be positive-definite and will yield direction of descent. In the other hand the following Proposition 7.3.2 shows that Wolfe's algorithm always yields steps such that $\langle \sigma_k, y_k \rangle > 0$ (if it is fed by a direction of descent)

Proposition 7.3.2 If d_{k-1} is a direction of descent, and if we use a Wolfe algorithm for the choice of step s_{k-1}, then $\langle \sigma_k, y_k \rangle > 0$, and B_k is positive definite, hence $d_k = -B_k \nabla f(x_k)$ is a direction of descent.

Proof

$$\begin{aligned}
\langle y_k, \sigma_k \rangle &= \langle \nabla f(x_k) - \nabla f(x_{k-1}), x_k - x_{k-1} \rangle \\
&= s_{k-1} \langle \nabla f(x_k) - \nabla f(x_{k-1}), d_{k-1} \rangle \\
&= s_{k-1} \langle \nabla f(x_k), d_{k-1} \rangle - s_{k-1} \langle \nabla f(x_{k-1}), d_{k-1} \rangle \\
&\geq s_{k-1} (\varepsilon_2 - 1) \langle \nabla f(x_{k-1}), d_{k-1} \rangle \text{ (2nd Wolfe rule)}, \\
&> 0,
\end{aligned}$$

since $\varepsilon_2 < 1$ and $\langle \nabla f(x_{k-1}), d_{k-1} \rangle < 0$.

So far, we have proven that BFGS algorithm is always possible and is a descent algorithm provided that it is paired with a Wolfe's linesearch. Since we have an exact formula for the inverse of H_k, there is no need to solve $H_k d_k = -\nabla f(x_k)$ to solve the linear system but we rather directly compute $d_k = B_k(-\nabla f(x_k))$, which is way faster. There is one last trick up the sleeves of the practitioner, do we compute B_k or is there something better we can do ? Suppose that we have a huge number of variables, say $n = 10^6$ (or $n = 1M$ for the picky readers), then storing B_k requires 10^{12} numbers, with 8 byte per number (this is the standard 64 bits encoding of floats), this yields to 8 TeraBytes of memory needed, which is out of reach. How do we solve this problem ? it turns out that a careful inspection of the update formula of B_k :

$$B_k = \left(I - \frac{\sigma_k y_k^T}{\langle y_k, \sigma_k \rangle} \right) B_{k-1} \left(I - \frac{y_k \sigma_k^T}{\langle y_k, \sigma_k \rangle} \right) + \frac{\sigma_k \sigma_k^T}{\langle y_k, \sigma_k \rangle},$$

reveals the recursive structure of the computation of $B_k r$, it is given here in pseudo code, we suppose that $(\sigma_k)_k$ and $(y_k)_k$ are stored in arrays denoted `sigma` and `y` respectively. We suppose here that $B_1 = Id$.

```
def ComputeB(k,f) :
  if k==1 :
    return f
   else
    tmp=sigma[k,:]*np.dot(sigma[k,:],f)/np.dot(y[k,:],sigma[k,:])
    f=f-y[k,:]*np.dot(sigma[k,:],f)/np.dot(y[k,:],sigma[k,:])
    f=computeB(k-1,f)
    f=f-sigma[k,:]*np.dot(y[k,:],f)/np.dot(y[k,:],sigma[k,:])
    return f+tmp
```

We see here that we do not need to store the whole matrix B_k, but rather the vectors $(\sigma_k)_k$ and $(y_k)_k$. It is an exercise to show that the BFGS algorithm can be rewritten the following way:

Proposition 7.3.3 Recall that $\sigma_k = x_k - x_{k-1}$ and $y_k = \nabla f(x_k) - \nabla f(x_{k-1})$. Denote $\rho_k = \frac{1}{\langle \sigma_k, y_k \rangle}$. And suppose that $(\sigma_k, y_k, \rho_k)_k$ is saved in a sequence denoted L. The descent direction of BFGS is given by the following algorithm :

```
1. $q = -\nabla f(x_k)$ and create an empty list called $L_\alpha$
2. For $(\sigma, y, \rho)$ in reversed order of $L$ :
    (a) Compute $\alpha = \rho\langle \sigma, q \rangle$ and append $\alpha$ to $L_\alpha$
    (b) Set $q = q - \alpha y$
3. Reverse the list of $L_\alpha$.
4. Set $q = B_0 q$.
5. For $(\sigma, y, \rho), \alpha$ in $(L, L_\alpha)$ :
    (a) Compute $\beta = \rho\langle y, q \rangle$.
    (b) Set $q = q + (\alpha - \beta)\sigma$
```

CHAPTER 8

Constrained smooth optimization

In this chapter, we describe several algorithms that can be used to tackle the problems with constraints. There are literally tons f algorithms, each one being competitive in its own area of expertise. It is safe to assume that there are as many algorithm as there are problems. By establishing that we can only scrap the surface of the possible algorithm, we are now free to begin our tour.

8.1 Projected gradient

The first algorithm we propose is restricted to the case where the set of constraints X is convex. We recall the theorem of projection on a convex set, proved in Theorem 2.3.3, page 31.

> Suppose that X is a closed convex set. For every x, solve $\min_{y\in X} \|y-x\|^2$. There is a unique solution denoted $\pi_X(x)$. This solution solves
>
> $$\langle \pi_X(x) - x, y - \pi_X(x)\rangle \geq 0 \quad \forall y \in K$$

We suppose that the set X is simple so that it is easy to compute π_X. The idea is to start with a gradient descent algorithm

$$\tilde{x}_{k+1} = x_k - s_k \nabla f(x_k).$$

Sadly the term $\tilde{x}_{k+1}$ might not belong to X, we might have exited X, the idea is then to project $\tilde{x}_{k+1}$ onto X in order to obtain x_{k+1}, that is

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

> **Definition 8.1.1 — Projected gradient algorithm.** Let E be a Hilbert space and $X \subset E$ a closed convex set, the goal is to minimize a function f over X. At each iteration k, choose a step s_k and perform the iteration
>
> $$x_{k+1} = \pi_{\mathcal{X}}(x_k - s_k \nabla f(x_k)).$$
>
> - If $d_k = x_{k+1} - x_k \neq 0$, it is a direction of descent of f at point x_k.

- For every $0 \leq \alpha \leq 1$, we have $x_k + \alpha d_k \in X$. The projected gradient algorithm is the choice $\alpha = 1$, but if necessary, α can be decreased (in a step reduction for the line search).

Proof

The points x_k and $x_{k+1} = x_k + d_k$ are in X which is convex. So that $x_k + \alpha d_k \in \mathcal{X}$ for every $0 \leq \alpha \leq 1$. We now want to prove that d_k is a direction of descent of f. Recall that, for every $y \in X$, we have :

$$\langle \pi_{\mathcal{X}}(x) - x, y - \pi_X(x) \rangle \geq 0.$$

Replace in the above inequality x by $x_k - s_k \nabla f(x_k)$ and y by x_k, then $\pi_{\mathcal{X}}(x) = x_{k+1}$ and we have

$$\langle x_{k+1} - (x_k - s_k \nabla f(x_k)), x_k - x_{k+1} \rangle \geq 0 \Longrightarrow \langle d_k + s_k \nabla f(x_k), -d_k \rangle \geq 0$$

$$\langle \nabla f(x_k), d_k \rangle \leq -\frac{1}{s_k} \|d_k\|^2 \leq 0$$

The projected gradient algorithm is an algorithm that yields a direction of descent. Moreover it has the same properties of convergence than the standard gradient algorithm (See Proposition **??**, page **??**). Of course $\nabla f(x_k)$ will not converge to zero, because it must verify KKT equation. Instead the direction $d_k = x_{k+1} - x_k$ willl converge to zero.

Proposition 8.1.1 If f has a L-Lipschitz gradient and is bounded from below and $\mathcal{X}$ is convex, then the fixed step projected gradient algorithm with $0 < s < \frac{2}{L}$ converges in the sense that $\sum \|x_{k+1} - x_k\|^2 < +\infty$.

Proof

Start with $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2$ and remember that if $d_k = x_{k+1} - x_k$, we have

$$\langle \nabla f(x_k), d_k \rangle \leq -\frac{1}{s} \|d_k\|^2 \leq 0.$$

Hence : $f(x_{k+1}) \leq f(x_k) + (\frac{L}{2} - \frac{1}{s}) \|d_k\|^2$ and proceed as usual (see Proposition **??**, page **??**).

8.2 Penalization method

The penalization method amounts to replace the minimization of f over X by the unconstrained minimization of an other function $f_\varepsilon = f + \varepsilon^{-1} \pi_X$. The function π_X must verify two properties

$$\pi_X(x) = 0 \text{ for } x \in X \text{ and } \pi_X(x) > 0 \text{ if } x \notin X.$$

The idea is to minimize f_ε for very small values of ε. Denote $x_\varepsilon^\star$ a minimum of f_ε, it might be outside X but the function π_X **penalizes** the fact that $x_\varepsilon^\star \notin X$. The hope is that, as ε goes to 0, then $x_\varepsilon^\star$ converges to $x^\star$ a minimum of f and that $x^\star$ belongs to x. In this section, we only deal with the simplest penalization possible, that is

Definition 8.2.1 — Definition of penalization. Suppose that $X \subset \mathbb{R}^n$, the penalization method amounts to replace the problem $\min_{x\in X} f(x)$ by

$$\min_{x\in\mathbb{R}^n} f_\varepsilon(x), \text{ with } f_\varepsilon(x) = f(x) + \frac{1}{2\varepsilon}\sum_i \tilde{g}_i^2(x),$$

and $\tilde{g}_i = g_i$ forall $i \in \mathcal{E}$ and $\tilde{g}_i = \max(g_i, 0)$ forall $i \in \mathcal{I}$.

It is easy to check that $X = \{x \text{ s.t. } \tilde{g}_i(x) = 0 \ \forall i\}$ and that forall i, we have

$$\frac{1}{2}\nabla(\tilde{g}_i^2)(x) = \tilde{g}_i(x)\nabla g_i(x).$$

We first begin by showing that the penalization defined above method converges

Proposition 8.2.1 — Convergence of the minimizers. With the notation of 8.2.1, if f is lower semi-continuous, coercive on $\mathbb{R}^n$ and bounded from below. If g_i is continuous for every i. Then let x_ε be a global minimizer of f_ε. Then there exists $x^\star$, a global minimizer of f on X such that x_ε converges up to a subsequence to $x^\star$ as $\varepsilon \to 0$.

Proof

The existence of a $\bar{x}$, a global minimizer of f on X is standard, remark that X is closed. The existence of x_ε follows from lower-semi continuity of f_ε and the fact that f_ε is coercive on $\mathbb{R}^n$. We have

$$\inf_{\mathbb{R}^n} f \le f(x_\varepsilon) \le f_\varepsilon(x_\varepsilon) \le f_\varepsilon(\bar{x}) = f(\bar{x}) = \inf_X f \tag{8.1}$$

Hence x_ε is bounded (because f is infinite at infinite), hence converges up to a subsequence to some $x^\star$. Denote $\pi_X = \sum_i \tilde{g}_i^2$, we have

$$0 \le \pi_X(x_\varepsilon) \le \varepsilon(f_\varepsilon(x_\varepsilon) - f(x_\varepsilon)) \underbrace{\le}_{(8.1)} \varepsilon(\inf_X f - \inf_{\mathbb{R}^n} f).$$

Hence we have $\pi_X(x_\varepsilon) = \mathcal{O}(\varepsilon)$ and as $\varepsilon \to 0$, we obtain $\pi(x^\star) = 0$, so that $x^\star \in X$. From (8.1), we retrieve $f(x_\varepsilon) \le \inf_X f$ and as ε goes to zero we obtain $f(x^\star) \le \inf_X f$ says that $x^\star$ is a global minimizer.

Proposition 8.2.1 is nice, but rather useless in practice. Indeed, Fixing ε using a unconstrained minimization algorithm to find x_ε the global minimizer of f_ε is an impossible task ! Note that, at best, we can hope to find a point such that $\nabla f_\varepsilon(x_\varepsilon)$ is small and we might possibly check that we are close to a **local minimizer**, but we cannot hope more. The next proposition is an answer to this problem

Proposition 8.2.2 Suppose that we are given $(\epsilon_k)_k$ and $(\eta_k)_k$ two non-increasing sequences of real positive number that converges to 0, suppose that there exists $M > 0$ and $(x_k)_k$ a sequence of vectors of $\mathbb{R}^n$ that verify :

$$\|\nabla f_{\varepsilon_k}(x_k)\| \le \eta_k \text{ and } f_{\varepsilon_k}(x_k) < M.$$

Such a sequence of (x_k) can be obtained by launching a unconstrained optimization algorithm on f_{ε_k} that starts at x_{k-1}. Then suppose that f and g_i are C^1, suppose that the sequence (x_k) is bounded, extract a subsequence and hence suppose that

$(x_k)_k$ converges to some $x^\star$. Suppose that the LICQ condition holds at $x^\star$, and denote for each i

$$\lambda_k[i] = \varepsilon_k^{-1}\tilde{g}_i(x_k).$$

Then $\lim_{k\to+\infty} \lambda_k = \lambda^\star$ exists and $(x^\star, \lambda^\star)$ is a KKT point.

One hypothesis is technical, and quite strange, the existence of M such that $f_{\varepsilon_k}(x_k) < M$ for every k. In order to make sure that such an M exists, it is sufficient to suppose that X is non-empty and to take any $z \in X$, then for each ε_k, $f_{\varepsilon_k}(z) = f(z)$. At iteration k, it is then sufficient to check wether $f_{\varepsilon_k}(x_{k-1}) \leq f(z)$ and to take the initial point with smallest initial value. If we implement a descent algorithm, we are sure that for each k, $f_{\varepsilon_k}(x_k) < M$ for $M = f(z)$.

Proof

- Suppose that $x^\star \notin X$, then $\pi_X(x^\star) = \frac{1}{2}\sum_i \tilde{g}_i(x^\star)^2 > 0$ and by continuity of π_X, we have $\pi_X(x_k) > \pi_X(x^\star)/2$ for large k. In this case we have for large k:

$$f_{\varepsilon_k}(x_k) \geq \inf_{\mathbb{R}^n} f + \varepsilon^{-1}\pi_X(x^\star)$$

  Hence $f_{\varepsilon_k}(x_k)$ tends to $+\infty$ which is in contradiction with the fact that it is bounded by M.
- We establish the convergence of λ_k.
  1. First if $i \in \mathcal{I} \setminus \mathcal{A}(x^\star)$, then $g_i(x^\star) < 0$ and for large k, we have $g_i(x_k) < 0$, hence $\tilde{g}_i(x_k) = 0$ and then $\lambda_k[i] = 0$. We proved that if $i \in \mathcal{I} \setminus \mathcal{A}(x^\star)$, $\lambda_k[i]$ converges to 0.
  2. Because LICQ condition is verified at $x^\star$, the family $(\nabla g_i(x^\star))_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}}$ is independent. Denote $(z_j)_j$ an independent family such that $(\nabla g_i(x^\star))_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}} \cup (z_j)_j$ is a basis. For each k, denote A_k the matrix whose columns are the vectors $(\nabla g_i(x_k))_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}} \cup (z_j)_j$. As k goes to $+\infty$, A_k converges to an invertible matrix $A^\star$. Since the determinant is a continuous function, then $det(A_k) \neq 0$ for large k, and hence A_k is invertible for large k.
  3. We have

$$\nabla f_{\varepsilon_k}(x_k) = \nabla f(x_k) + \varepsilon_k^{-1}\sum_i \tilde{g}_i(x_k)\nabla g_i(x_k).$$

  And hence

$$\sum_i \lambda_i[k]\nabla g_i(x_k) = \nabla f_{\varepsilon_k}(x_k) - \nabla f(x_k)$$

  Set k large enough so that $\lambda_i[k] = 0$ for all $i \in \mathcal{I}\setminus\mathcal{A}_{x^\star}$. Denote γ_k a vector in $\mathbb{R}^n$ such that its first coordinates are given by $(\lambda_k[i])_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}}$ and the rest is zero. We then have

$$A_k\gamma_k = \nabla f_{\varepsilon_k}(x_k) - \nabla f(x_k) \Rightarrow \gamma_k = A_k^{-1}(\nabla f_{\varepsilon_k}(x_k) - \nabla f(x_k))$$

  As k goes to $+\infty$, the vector γ_k converges to some vector $\gamma^\star = (A^\star)^{-1}(-\nabla f(x^\star))$. Denote $(\lambda^\star[i])_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}}$ the first coordinates of the vector $\gamma^\star$. Trivially the rest of the coordinates of $\gamma^\star$ are 0. We have convergence of $\lambda_k[i]$ towards $\lambda^\star[i]$ for all i and the equation

$$A^\star\gamma^\star + \nabla f(x^\star) = 0,$$

is exactly

$$\sum_{i\in\mathcal{A}_{x^\star}\cup\mathcal{E}} \lambda_i^\star \nabla g_i(x^\star) + \nabla f(x^\star) = 0,$$

which means that $x^\star$ is a KKT point of f on X.

8.3 Barrier method

The barrier method is very similar to the penalization, it amounts to replace the minimization of f over X by the unconstrained minimization of an other function $f_\varepsilon = f + \varepsilon\pi_X$. The function π_X must verify two properties

$$\pi_X(x) = +\infty \text{ for } x \in X \text{ and } \pi_X(x) \geq 0 \text{ if } x \in X \text{ and } \pi_X \text{ is regular .}$$

The idea is to minimize f_ε for very small values of ε. Denote $x_\varepsilon^\star$ a minimum of f_ε, it cannot escape X and the function π_X acts as a **barrier** that prevents x to exit X. The hope is that, as ε goes to 0, the barrier vanishes and $x_\varepsilon^\star$ converges to $x^\star$ a minimum of f. Note that it is not possible to implement barrier functions for equality function.

Definition 8.3.1 — Definition of Barrier method. Suppose that $X \subset \mathbb{R}^n$ is defined by inequalities only, the barrier method amounts to replace the problem $\min_{x\in X} f(x)$ by

$$\min_{x\in\mathbb{R}^n} f_\varepsilon(x), \text{ with } f_\varepsilon(x) = f(x) - \varepsilon\sum_{i\in\mathcal{I}} \log(-g_i(x)),$$

It is easy to check that the domain of f_ε is included in $X = \{x \text{ s.t. } g_i(x) \leq 0\ \forall i \in I\}$ and that forall i. Note that it is not possible to put define barriers for equality constraint. If one wants to tackle a problem with equalities and inequalities and have barriers for the inequalities, he still has to deal with equalities with a penalization method. We first begin by showing that the barrier method converges

Proposition 8.3.1 — Convergence of the minimizers. With the notation of 8.3.1, Suppose that

- f is lower semi-continuous, coercive on X and bounded from below.
- $X^o = \{x \text{ s.t } g_i(x) < 0\}$ is non-empty.
- g_i is continuous for every i

Then for all ε, there x_ε be a global minimizer of f_ε and the sequence $(x_\varepsilon)_\varepsilon$ converges up to a subsequence to some $x^\star \in X$ as $\varepsilon \to 0$.

- If LICQ conditions are verified for one global minimizer of f, then $x^\star$ is a global minimizer of f on X.
- If in addition LICQ conditions are verified at $x^\star$. Denote $\lambda_\varepsilon[i] = \frac{\varepsilon}{-g_i(x_\varepsilon)}$ then λ_ε converges to $\lambda^\star$ such that $(x^\star, \lambda^\star)$ is a KKT point.

The proof of the convergence of barrier functions is not done in main lecture, it is left as an exercise (with corrections) to the reader.

Exercice 8.1

1. Show that the problems $\inf_X f_\varepsilon$ and $\inf_X f$ admits global minimum.
2. Show that for all $x \in X$, we have :

$$inf_X f \leq f(x_\varepsilon) \leq f_\varepsilon(x_\varepsilon) \leq f_\varepsilon(x) \leq f_1(x) \tag{8.2}$$

3. Show that the sequence $(x_\varepsilon)_\varepsilon$ converges up to a subsequence to some $x^\star \in X$ as ε goes to 0.
4. We prove that $x^\star$ is a global minimum. Take $\bar{x}$ a global minimizer of f for which LICQ conditions are verified.
   (a) Show that there exists a direction u such that $(\nabla g_i(\bar{x}), u) < 0$ for each $i \in A_{\bar{x}}$. Conclude that for every $\alpha > 0$ small enough if $z_\alpha = \bar{x} + \alpha u$, then $g_i(z_\alpha) < 0$ for each i.
   (b) Let $\eta > 0$. Show that there exists $\varepsilon > 0$ small enough such that $f(z_\alpha) < f(\bar{x}) + \eta$.
   (c) Conclude that $f(x^\star) \leq f(\bar{x}) + \eta$.
   (d) Show that $x^\star$ is a global minimizer of f on X.
5. We prove that λ_ε converges to $\lambda^\star$
   (a) Show that
   $$\nabla f(x_\varepsilon) + \sum_i \lambda_\varepsilon[i] \nabla g(x_\varepsilon) = 0$$
   .
   (b) Suppose that $g_i(x^\star) < 0$, show in this case that $\lambda_\varepsilon[i]$ converges to $\lambda_i^\star$.
   (c) Conclude on the convergence of λ_ε towards $\lambda^\star$

Solution to Exercice 8.1

1. It is standard to show that f admits a minimum over X. We have $f_\varepsilon \geq f$, hence f_ε is coercive on X. Moreover there exists a point in X^o, hence $\inf_X f_\varepsilon \neq +\infty$. Hence f_ε admits a global minimum on X. Now $f_\varepsilon = +\infty$ outside X, hence minimzing f_ε on $\mathbb{R}^n$ amounts to minimizing it on X.
2. The inequalities are quite trivial to show.
3. We first state that the norm of $(x_\varepsilon)_\varepsilon$ is bounded. If this is not the case, there exists a subsequence such that $\|x_\varepsilon\|$ converges to $+\infty$. But in this case $f(x_\varepsilon)$ must converge to $+\infty$, since f is coercive. By (8.2), this shows that for every $x \in X^o$ $f_1(x) = +\infty$ which is absurd.
4. (a) Let $e_i = \nabla g_i(\bar{x})$ and $E = Vect(e_i)_{i \in \mathcal{A}_{\bar{x}}}$ By the LICQ condition, $(e_i)_{i \in \mathcal{A}_{\bar{x}}}$ is a basis of E,
      - We prove that there exists a vector such that $(e_i, u) < 0$ for each i. To do so, construct the matrix $A_{ij} = (e_i, e_j)$ it is a square invertible matrix. Indeed if x is such that $Ax = 0$, then the vector $\sum x_i e_i$ is orthogonal to every e_j and belongs to E. Hence $\sum x_i e_i$ must be equal to 0 which means that $x = 0$ because the e_i are linearly independent. We then find x such that $Ax = -b$, with b a vector with 1 for every coordinate. We then have $u = \sum x_i e_i$ verifies that $(e_i, u) = -1$ for each i.
      - For small α we have
      $$g_i(\bar{x} + \alpha u) = g(\bar{x}) + \alpha(e_i, u) + \mathcal{O}(\alpha)$$
      Hence $g_i(\bar{x} + \alpha u) < 0$ for small enough α and every $i \in \mathcal{A}_{\bar{x}}$. If $i \notin \mathcal{A}_{\bar{x}}$ because $g(\bar{x}) < 0$ and continuity of g_i, we also have $g_i(\bar{x} + \alpha u) < 0$ for small enough α.

   (b) Let $\eta > 0$. For α small enough, by continuity of f, we have for α small enough $f(x_\alpha) < f(\bar{x}) + \eta$.
   (c) Take $\eta > 0$ and α small enough such that $f(z_\alpha) < f(\bar{x}) + \eta$, upon taking

α even smaller, we have $g_i(z_\alpha) < 0$ for all i, hence $\pi(x_\alpha) \neq +\infty$. We then have

$$f(x_\varepsilon) \leq f_\varepsilon(x_\alpha) \leq f(\bar{x}) + \eta + \varepsilon\pi_X(x_\alpha).$$

Let ε goes to 0, then $f(x^\star) \leq f(\bar{x}) + \eta$.

(d) Since η is arbitrary, then $f(x^\star) \leq f(\bar{x})$. Use $f(\bar{x}) = \inf_X f$ and its done.

5. (a) We have $\nabla f_\varepsilon(x_\varepsilon) = 0$. Hence

$$\nabla f_\varepsilon + \varepsilon \nabla \pi_X(x_\varepsilon) = 0.$$

which gives

$$\nabla f_\varepsilon + \sum_i \lambda_\varepsilon[i] \nabla g(x_\varepsilon) = 0 \text{ if } \lambda_\varepsilon[i] = \frac{\varepsilon}{-g_i(x_\varepsilon)}.$$

(b) If $g_i(x^\star) < 0$ then $g_i(\varepsilon) < \frac{g_i(x^\star)}{2}$ for ε small enough and $\frac{\varepsilon}{-g_i(x_\varepsilon)}$ converges to 0. Incidently $\lambda_i^\star = 0$ in this case.

(c) Denote $u_\varepsilon = \nabla f(x_\varepsilon) + \sum_{i \in \mathcal{A}_{x^\star}} (\lambda_\varepsilon)_i \nabla g_i(x_\varepsilon)$, we have that u_ε converges to $\nabla f(x^\star)$ as ε goes to zero and

$$u_\varepsilon + \sum_{i \in \mathcal{A}_{x^\star}} (\lambda_\varepsilon)_i \nabla g_i(x_\varepsilon) = 0$$

As ε goes to 0 we have

$$\lim_{\varepsilon \to 0} u_\varepsilon + \sum_{i \in \mathcal{A}_{x^\star}} \lambda_i^\star \lim_{\varepsilon \to 0} \nabla g_i(x_\varepsilon) = 0$$

The fact that each $(\lambda_\varepsilon)_i$ converges to $(\lambda^\star)_i$ comes from the fact that the family $(\nabla g_i(x^\star))_{i \in \mathcal{A}_{x^\star}}$ is linearly independent and hence the decomposition of $\nabla f(x^\star)$ on this family is unique and given by

$$\nabla f(x^\star) = -\sum_{i \in \mathcal{A}_{x^\star}} \lambda_i^\star \nabla g_i(x^\star)$$

INSA TOULOUSE
135 avenue de Rangueil
31400 Toulouse

Tél : + 33 (0)5 61 55 95 13
www.insa-toulouse.fr

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE