INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Optimization

3MIC et 4MA

Dr. Frédéric de Gournay
Dr. Aude Rondepierre

Publié par INSA de Toulouse

degourna@insa-toulouse.fr
rondepierre@insa-toulouse.fr

Première édition, Septembre 2023

Contents

Introduction

CHAPTER 1

Presentation

♣ 1.1 Presentation of optimization

Optimization and in particular numerical optimization has undergone a tremendous boom for the past 50 years due to the ever increasing computational power.

CHAPTER 2

Refresher

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

♣ 2.1 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 2.1.1 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 :

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

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 2.1.2 — 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**.

Exercise 2.1

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 of Exercise 2.1

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

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 2.1.3 — 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:

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

Exercise 2.2

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 of Exercise 2.2

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

Theorem 2.1.1 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 2.1.4 — 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:

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

Exercise 2.3

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

Solution of Exercise 2.3

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

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

Theorem 2.1.2 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 2.1.5 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

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

Exercise 2.4

Compute the gradient and the Hessian of

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

Solution of Exercise 2.4

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

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 2.1.3 — 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.

♣ 2.2 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 2.2.1 — Classic Taylor expansions. If $f : \mathbb{R} \mapsto \mathbb{R}$ is sufficiently regular, we have:

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

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

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

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

Proposition 2.2.2 — 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 2.2.3 — Real valued function. If $f : \mathbb{R}^n \mapsto \mathbb{R}$, is regular enough, we have :

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

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

♣ 2.3 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 2.2, 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 2.2. The vector $\nabla f(x_0, y_0)$ attached at the point (x_0, y_0) and is displayed in red in Figure 2.2.

♣ 2.4 Symetric matrices

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

Definition 2.4.1 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

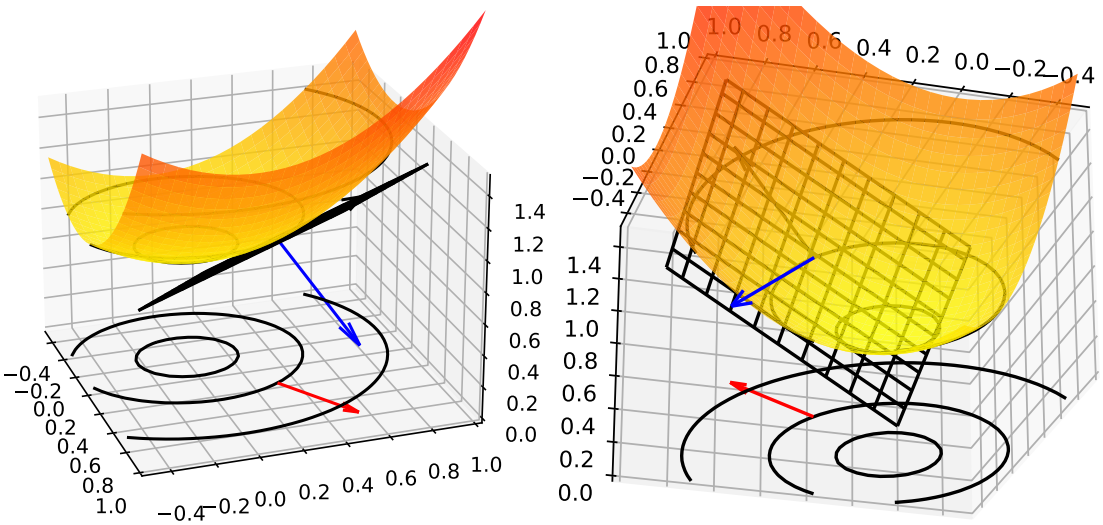


Figure 2.1: A quadratic function and its Taylor expansion in $(x_0, y_0) = (0.4, 0.3)$ its levelset are in black and its tangent plane at (x_0, y_0) is in black. In red, the gradient at (x_0, y_0) , in blue the vector $\begin{pmatrix} \nabla f(x_0, y_0) \\ -1 \end{pmatrix}$, in black the vector $\begin{pmatrix} \nabla f(x_0, y_0) \\ \|\nabla f(x_0, y_0)\|^2 \end{pmatrix}$.

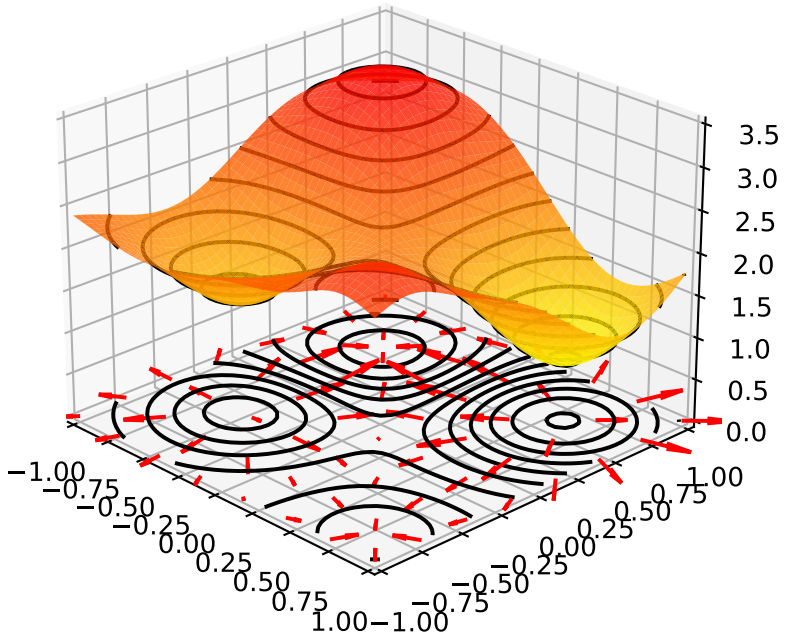


Figure 2.2: A function, its levelsets and the gradient field. Notice that the gradient is orthogonal to the level-sets.

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 2.4.1 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

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

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).

CHAPTER 3

Starters

♣ 3.1 Partial order

The notation $a \succeq b$ or $a \succ b$ depends on the type of a and b. In order to make the notation clear, we list the three main notations of $\succeq$ or $\succ$, these notations are linked to a certain cone, we give the name of the cone.

Definition 3.1.1 The definition $a \succeq 0$ or $a \succ 0$ depends on the type of a.

- If $a \in \mathbb{R}$, then $a \succeq 0$ iff $a \geq 0$ and $a \succ 0$ iff $a > 0$. This is the usual definition of the order in $\mathbb{R}$.
- If $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, then $a \succeq 0$ iff $a_i \geq 0$ for every i and $a \succ 0$ iff $a_i > 0$ for every i. This is the partial order associated to the positive orthant.
- If a is a symetric matrix, we saw that $a \succeq 0$ iff every eigenvalue of a is non-negative and $a \succ 0$ iff the eigenvalues are positive. This is the partial order associated to the cone of positive matrices.

When the definition $a \succ 0$ is given, we can define $a \succ b$ and $a \prec b$ in the following manner.

Definition 3.1.2 We say that $a \succ b$ if and only if $a - b \succ 0$. We say that $a \prec b$ if and only if $b \succ a$. There are equivalent definitions for $\succeq$ and $\preceq$.

Exercise 3.1

Let A be a square matrix and $\lambda \in \mathbb{R}$, prove that $A \succeq \lambda \mathrm{Id}$ iff every eigenvalue of A is greater than or equal to λ.

Solution of Exercise 3.1

Denote $\mathrm{Sp}(A)$ the set of eigenvalues of A. Then $\mathrm{Sp}(A - \lambda \mathrm{Id}) = \{\mu - \lambda \text{ for } \mu \in \mathrm{Sp}(A)\}$. Then $A \succeq \lambda \mathrm{Id}$ iff $A - \lambda \mathrm{Id} \succeq 0$, which means that every eigenvalue of $A - \lambda \mathrm{Id}$ is positive, which is equivalent to saying that every egeinvalue of A

must be $\geq \lambda$.

The important thing to remember about these partial orders is that they behave well with respect to the usual operations but that we cannot always compare two objects.

Proposition 3.1.1 — Usual operations of partial orders.
- If $a \succ b$ and $c \succ d$ then $a + c \succ b + d$.
- If $a \succ b$ and $\lambda > 0$ then $\lambda a \succ \lambda b$.
- If $a \succ b$ and $\lambda < 0$ then $\lambda a \prec \lambda b$.
- There exists examples where neither $a \succ b$ nor $a \prec b$.

Proof

Only the last item is important to remember. Take for instance $a = (-1, 1)$ and $b = (0, 0)$. It is not true that $a \succ 0$ and neither do we have $a \prec b$. An other example is when a is a symetric matrix with one positive eigenvalue and one negative eigenvalue, then $a \not\succ 0$ and $a \not\prec 0$.

♣ 3.2 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 : $f(x + h) = f(x) + f'(x)h + \mathcal{O}(h^2)$

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

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

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

Proposition 3.2.1 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 $y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{3.1}$$

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 (3.1) and the Cauchy-Schwartz inequality, indeed we have:

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

we obtain

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

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

Proposition 3.2.2 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

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

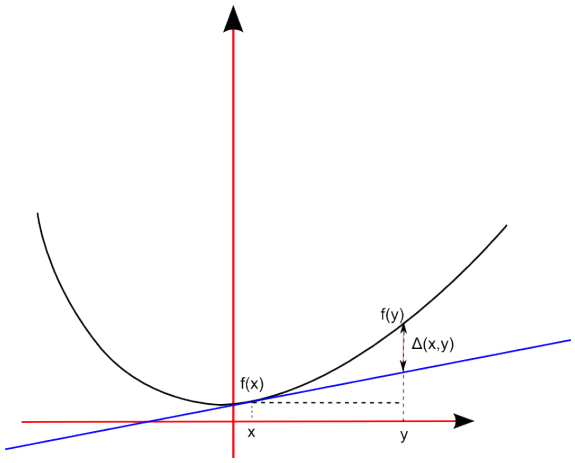


Figure 3.1: An example of function with Lipschitz gradient, we represent the function, its tangent and the two parabolas $y \mapsto f(x) + \langle \nabla f(x), y - x \rangle \pm \frac{L}{2}\|y - x\|^2$. The function is allowed to take any value in-between the two parabolas. It cannot stray too far apart from the first order approximation.

Exercise 3.2

1. The function $f(x) = \frac{1}{2}\|Ax - b\|^2$ has Lipschitz gradient of constant $L = \lambda_{max}(A^T A)$.
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.

♣ 3.3 Exercise

♣ 3.3.1 Some exercises

♠ 3.3.2 More exercises

Exercise 3.3

Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and $b \in \mathbb{R}^m$.

1. Let
$$\Psi(y) = \sum_{i=1}^{m} y_i^4 + 1 \quad \text{and} \quad f(x) = \Psi(Ax).$$
Compute $\nabla\Psi(y)$ and $H[\Psi]$. Is Ψ convex ?
2. Let $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ be defined by $\varphi(x) = Ax - b$. Compute the jacobian of φ
3. Let $f(x) = \Psi(Ax - b)$ compute ∇f by two different methods.

Exercise 3.4

Compute the gradient and the hessian of the mapping

$$f : x \in \mathbb{R}^n \mapsto \tfrac{1}{2}\|Ax - b\|^2 \text{ where } A \in \mathcal{M}_{m,n}(\mathbb{R}) \text{ and } b \in \mathbb{R}^m.$$

Under which condition on A is the function f convex ? strictly convex ?

Exercise 3.5

We recall that $\mathcal{M}_{n,n}(\mathbb{R})$, the vector space of real square matrices of size n is Euclidean when endowed with the scalar product $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. Show quickly that each of the following map is differentiable on its domain, compute their differential and their gradient.

1. $\mathrm{tr} : A \mapsto \mathrm{tr}(A)$
2. $\det : A \mapsto \det(A)$ when A is invertible (hint, first prove that $\nabla \det(I_n) = I_n$).
3. $f : A \mapsto \ln|\det(A)|$ when A is invertible

Exercise 3.6

In the space $\mathcal{M}_{m,n}(\mathbb{R})$ of real matrices of size $m \times n$ with the scalar product $\langle U, V\rangle = \mathrm{tr}(U^\top V)$. Let $B \in \mathcal{S}_n$ a symetric $n \times n$ real matrix. Let

$$f : A \in \mathcal{M}_{m,n}(\mathbb{R}) \mapsto tr(ABA^\top)$$

Give the differential and the gradient of f

II Convexity

CHAPTER 4

Definitions and basic properties

The notion of convexity is fundamental to both the theory and practice of modern optimization. Hence we dedicate a whole chapter to its study. There are two different notions of convexity, the convexity of a function and the convexity of a set. We start by the latter.

♣ 4.1 Convexity of sets

♣ 4.1.1 Definition

Definition 4.1.1 — 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

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

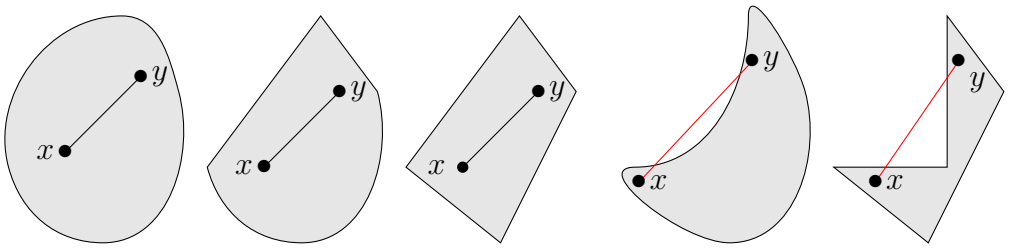


Figure 4.1: Convexity of sets, the first three sets are convex while the last two are not.

In words, a set $\mathcal{C}$ is convex if and only if it contains every segment that ends in $\mathcal{C}$. In Figure 4.1, we display several examples of convex sets. In Figure 4.2, we emphasize the fact that the boundary is of importance when considering convexity. Indeed, in Figure 4.2, we represent four sets which have same interior and same closure, only the first three of those sets are convex.

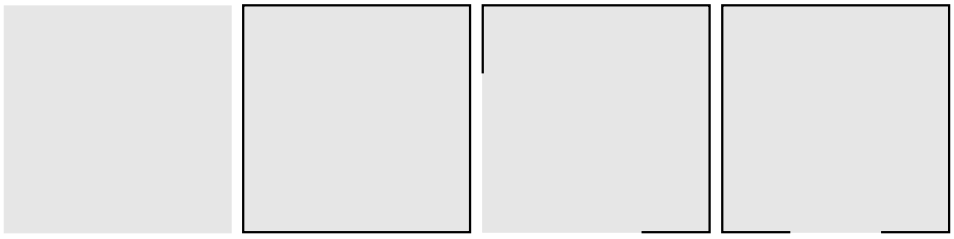


Figure 4.2: Convexity of sets, the first three sets are convex while the last one is not. In black is the part of the boundary that belongs to the set.

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.

Proposition 4.1.1 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 4.1.2 — 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**).

♡ 4.1.2 Convex Hull

Definition 4.1.2 — 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 4.3, we display some examples of family $(x_i)_{i\in I}$ and the corresponding convex combinations.

Definition 4.1.3 — Convex hull. Let $X \subseteq \mathbb{R}^n$, the convex hull of X is denoted $\text{conv}(X)$ 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 :

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

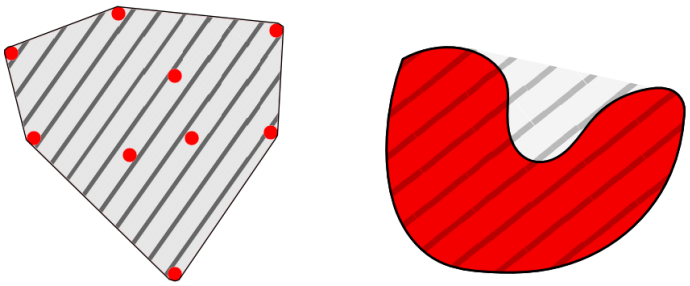


Figure 4.3: Exemples of convex hulls. On the left, a typical convex hull of a finite number of points. On the right, a convex hull of a domain with an infinite number of points.

♣ 4.2 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 4.2.1 — 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 4.4. 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.

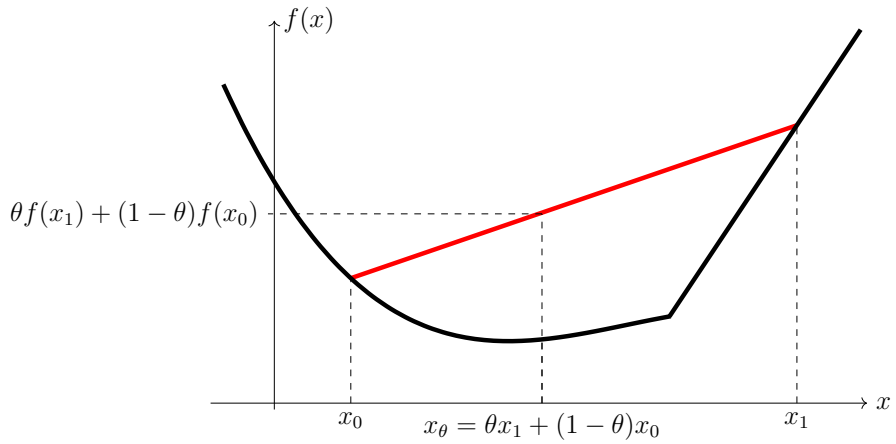


Figure 4.4: An example of convex function from $\mathbb{R}$ to $\mathbb{R}$. Note that this function is not differentiable everywhere

In Figure 4.4, the segment between $(x, f(x))$ and $(y, f(y))$ is called **a chord**. By Definition 4.2.1, 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 4.5 in $2d$ and $3d$.

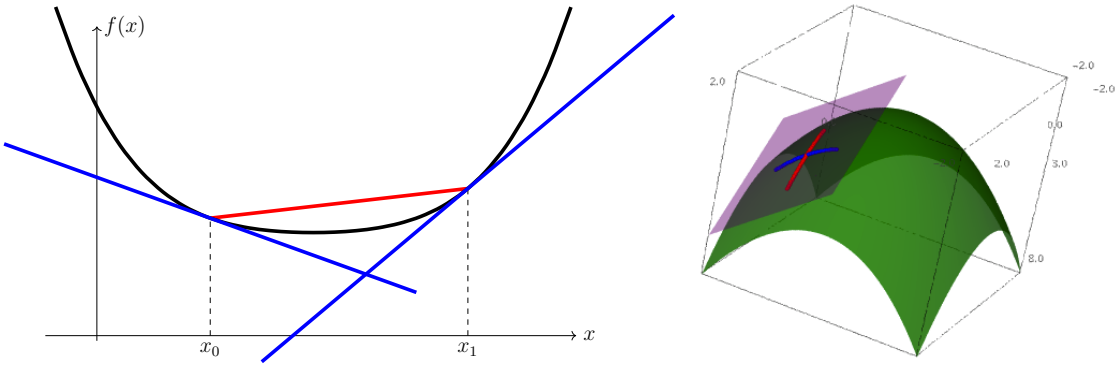


Figure 4.5: A convex function is always above its tangent hyperplanes (when they exists), on the left a $1d$ example and on the right a $2d$ example.

Theorem 4.2.1 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

$$\forall (x, y) \in X^2, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \tag{4.1}$$

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 (4.1). We now suppose that (4.1) is true. For any $(a, b) \in X^2$ and $\theta \in [0, 1]$, apply (4.1) to $(x = \theta a + (1-\theta)b, y = a)$, and then to $(x = \theta a + (1-\theta)b, y = b)$, we obtain

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

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 4.2.1 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 4.2.2 — 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

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

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 :

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

♣ 4.3 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 4.3.1 — 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 4.3.2 — 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 $x \in \mathring{X}$ (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 $x \in \mathring{X}$, then $x + \varepsilon d \in X$ for every $\varepsilon > 0$ small enough. Moreover

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

Hence if f is convex, there is a contradiction with Theorem 4.2.1.

Theorem 4.3.3 — 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 4.3.4 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 4.3.3

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 (**1**) as the first order derivative of ϕ and (**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 4.3.4 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.

♡ 4.4 Domain and epigraph

In this section, we study a little deeper the connection between a convex set and a convex function. First we see that any convex set can be turned into a convex function, by the use of the **indicatrix** and the **domain**. Second we see that studying convex functions amounts to study convex sets, the main tool of this analysis is the **epigraph**. This ability to juggle between convex functions and convex sets is the art of the master of convexity. It will soon become one of your favorite skill.

♡ 4.4.1 Domains

In optimization, we often allow functions to be equal to $+\infty$ at some points, so that we consider functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$.

There is a rule of thumb here, we allow the function to be equal to $+\infty$ but we **NEVER** allow it to be equal to $-\infty$. Hence, for a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, never write the function $-f$, or you are going into real problems. In the following propositions we list the authorized algebra :

Proposition 4.4.1 Let a and $b \in \mathbb{R} \cup \{+\infty\}$. Then the following operations always make sense

$$a + b \quad \text{and} \quad a = b \quad \text{and} \quad a \leq b \quad \text{and} \quad \lambda a \text{ if } \lambda > 0$$

The following operations do not have a meaning

$$a * (+\infty) \text{ when } a \leq 0 \quad \text{and} \quad \frac{a}{+\infty} \text{ when } a < 0$$

Morally, we can add functions and multiply them by stictly positive number and that's all ! The set of functions from $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is not a vector space, when it is encountered for the first time, it makes things difficult. But when the student realizes that this set of function is a convex cone, he is very happy. Allowing functions to be equal to $+\infty$ calls for the notion of domain of a function.

Definition 4.4.1 — Domain of a function. The domain of a function f with values in $\mathbb{R} \cup \{+\infty\}$ is denoted $dom(f)$ and is defined as the set of points where the function is finite-valued

$$dom(f) = \{x \in \mathbb{R}^n, f(x) < +\infty\}.$$

Autorizing functions to be equal to $+\infty$ allows us to get rid of the constraints. Indeed if we consider the problem $\inf_{x \in X} f(x)$ where X is a non-empty subset of $\mathbb{R}^n$ and if we define

$$\bar{f}(x) = \begin{cases} f(x) & \text{if } x \in X, \\ +\infty & \text{else.} \end{cases}$$

Then we have:

$$\inf_{x \in X} f(x) = \inf_{x \in \mathbb{R}^n} \bar{f}(x).$$

The interest of this trick is that in the convex realm, most of the theorem with constraints can be rephrased in the unconstrained case. The general setting is the following

Definition 4.4.2 For any set $X \subset \mathbb{R}^n$, define $\mathbb{1}_X$ the function

$$\mathbb{1}_X(x) = \begin{cases} 0 & \text{if } x \in X \\ +\infty & \text{if } x \notin X \end{cases}$$

For any function $f : \mathbb{R}^n \to \mathbb{R}$, define $\bar{f} = f + \mathbb{1}_X$, then $\bar{f}$ is a function from $\mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and minimizing f over X is the same problem than minimizing $\bar{f}$.

In what follows, we will always suppose that the functions under consideration have non-empty domain (i.e., we are not dealing with the function which is always equal to $+\infty$). We redefine the notion of convexity of a function

Definition 4.4.3 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We say that f is convex iff
1. The domain of f is a convex set in the sense of definition 4.1.1.
2. The function f is convex over its domain, in the sense of definition 4.2.1.

This definition can be conveniently be put into a simple proposition

Proposition 4.4.2 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We say that f is convex iff

$$\forall x, y \in \mathbb{R}^n, \text{ and } \theta \in [0,1], \ f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y). \tag{4.2}$$

Proof

Supppose that f is convex, then if x and y are in the domain of f, then because of definition 4.2.1 , then (4.2) is true. If x or y are not in $dom(f)$, then the right hand-side of (4.2) is $+\infty$, and hence (4.2) is true. Suppose now that (4.2) is true. Take x and y in $dom(f)$, then $\theta f(x) + (1-\theta) f(y) < +\infty$, and then $f(\theta x + (1-\theta)y) < +\infty$ and then $\theta x + (1-\theta)y \in dom(f)$. Hence $dom(f)$ is a convex set. Restricting x and y to belong to $dom(f)$ in (4.2) proves that f is convex over its domain.

As a byproduct of the above theorem, a set is convex iff its indicatrix is convex.

♡ 4.4.2 Epigraph

The indicatrix of a set allows us to transform the study of convexity of sets into the study of convexity of functions. This has a cost, because the indicatrix of a set is not a very nice function, it only takes the values 0 and $+\infty$. It turns out that there exists a way to transform the study of convexity of functions into the study of convexity of sets. The resulting object, called the **epigraph** is nicer than the indicatrix function.

Theorem 4.4.3 Let f be a function with values in $\mathbb{R} \cup \{+\infty\}$. Its epigraph is defined as

$$epi(f) = \{(x,t) \in dom(f) \times \mathbb{R},\ t \geq f(x)\}.$$

A function f is convex if and only if its epigraph a convex set.

Proof

Suppose that f is convex. Let $M_1 = (x_1, t_1) \in epi(f)$ and $M_2 = (x_2, t_2) \in epi(f)$ then for every $\theta \in [0,1]$, we have

$$\theta t_1 + (1-\theta)t_2 \geq \theta f(x_1) + (1-\theta) f(x_2) \geq f(\theta x_1 + (1-\theta)x_2).$$

Hence $\theta M_1 + (1-\theta)M_2 = (\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2) \in epi(f)$.
Suppose now that $epi(f)$ is convex, then for every x_1, x_2 in $dom(f)$, we have

$$(x_1, f(x_1)) \in epi(f),\ (x_2, f(x_2)) \in epi(f).$$

Hence, for every $\theta \in [0,1]$, $(\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2) \in epi(f)$, or equivalently

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

The epigraph of a function is the zone above the graph. The zone below the graph is called the **hypograph**, it is not interesting to study the hypograph of a function, unless one is interested in the study of concave function. Indeed the hypograph of f is exactly the epigraph of $-f$. The epigraph must not be confused with the **levelset** of a function.

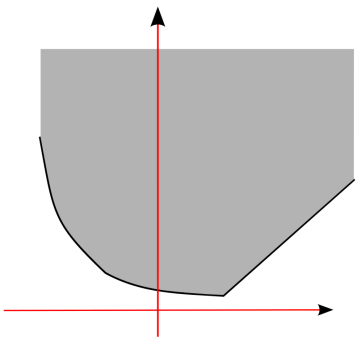


Figure 4.6: The epigraph of a function is the zone above the graph. Here the graph is in black and the epigraph in light gray.

Theorem 4.4.4 If f is convex, then its level-sets

$$\mathcal{L}_f(\beta) = \{x \in dom(f), f(x) \leq \beta\}$$

are either empty or convex.

Proof

If x_1 and x_2 belong to $\mathcal{L}_f(\beta)$, then $\forall \alpha \in [0,1]$,

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \leq \alpha\beta + (1-\alpha)\beta = \beta.$$

If the levelsets of a function are convex, there is no reason for a function to be convex. For instance, in 1D, take $f(x) = \sqrt{|x|}$. If every levelset of a function is convex, then the function is said to be **quasi-convex**. an example of such a function is given in 4.7. Note that a lot of theorem of convexity are overkills in the sense that the hypothesis f **is convex** can be replaced by f **is quasi-convex** and still be true. However, for simplicity, we will not study quasi-convexity in this book.

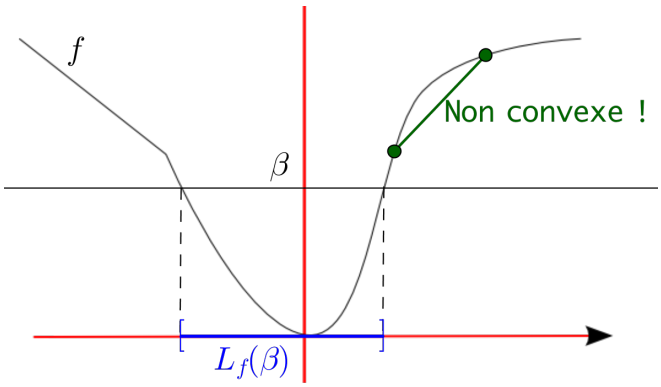


Figure 4.7: An example of quasi-convex function. The level-sets of this function are segments (hence convex sets) but the function is not convex..

Standard operations on convex functions

Theorem 4.4.5 Let f_1 and f_2 be two convex functions with values in $\mathbb{R} \cup \{+\infty\}$ and $\beta > 0$. The following functions are convex :

- $f(x) = \beta f_1(x)$.
- $f(x) = (f_1 + f_2)(x)$ and $dom(f) = dom(f_1) \cap dom(f_2)$.
- $f(x) = \max\{f_1(x), f_2(x)\}$ and $dom(f) = dom(f_1) \cap dom(f_2)$.

Theorem 4.4.6 Let $\phi : \mathbb{R}^m \to \mathbb{R}$ be a convex function. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator and $b \in \mathbb{R}^n$. Then the function $f : x \mapsto \phi(Ax + b)$ is convex and $dom(f) = \{x \in \mathbb{R}^n \mid Ax + b \in dom(\phi)\}$.

Proof

Let $(x_1, x_2) \in dom(f)$. Denote $y_1 = Ax_1 + b$ and $y_2 = Ax_2 + b$. Then for every

$\alpha \in [0,1]$:

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= \phi(\alpha(Ax_1+b) + (1-\alpha)(Ax_2+b)) \\ &\leq \alpha\phi(Ax_1+b) + (1-\alpha)\phi(Ax_2+b) \\ &\leq \alpha f(x_1) + (1-\alpha)f(x_2). \end{aligned}$$

♣ 4.5 Exercise

♣ 4.5.1 Some exercises

Exercise 4.1
Draw the following functions and prove that they are convex or prove that they are not

1. $f : x \mapsto x^2$
2. $f : x \mapsto e^{-x^2}$
3. $f : x \mapsto e^x$
4. $f : x \mapsto \sqrt{x}$
5. $f : x \mapsto e^{-x}$
6. $f : x \mapsto \frac{1}{x}$

Exercise 4.2
Discuss about the convexity or concavity of the following functions

1. $f(x,y) = (y-x^2)^2 - x^2$
2. $f(x,y) = x^4 + y^4 - (x-y)^2$
3. $f(x,y,z) = x^4 + 2y^2 + 3z^2 - yz - 23y + 4x - 5$

Exercise 4.3
Let E be a vector space with a norm $\| \bullet \|$, show that the norm is a convex function.

Exercise 4.4
Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a convex non-decreasing function and $h : \mathbb{R}^n \mapsto \mathbb{R}$ be a convex function. Show that $f = g \circ h$ is convex. Give an example where g and h are convex but $g \circ h$ is non-convex.

♠ 4.5.2 More exercises

Exercise 4.5

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be defined as $f(x) = \log(e^{x_1} + \dots e^{x_n})$. For any $x \in \mathbb{R}^n$, denote $s \in \mathbb{R}^n$ such that $s_i = \frac{e^{x_i}}{e^{x_1}+\dots+e^{x_n}}$.

1. Show that
$$\nabla f(x) = s \quad \text{and} \quad \begin{cases} (H[f](x))_{ii} = s_i^2 - s_i s_i \\ (H[f](x))_{ij} = -s_i s_j \qquad \text{if } i \neq j \end{cases}$$
2. Show that $(H[f](x)h, h) = \sum_i s_i h_i^2 - (\sum_i s_i h_i)^2$.
3. From $\sum s_i = 1$ and $s_i > 0$, conclude that f is (strictly) convex.

Exercise 4.6

Are the following sets convex ?

1. $X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1,\ y \geq x^2\}$
2. $X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \geq 1,\ y \geq x^2\}$
3. $X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1,\ y \leq x^2,\ y \leq 0\}$
4. $X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1,\ y = 2x\}$

Exercise 4.7

Let g be a concave positive function. Show that $f = \frac{1}{g}$ is convex

♡ 4.5.3 Even more exercises

Exercise 4.8

Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set, let λ_1 and λ_2 be two positives real number.

1. Draw the set $\lambda_1\mathcal{C} + \lambda_2\mathcal{C} = \{\lambda_1 x + \lambda_2 y \text{ such that } x, y \in \mathcal{C}\}$.
2. Show that $\lambda_1\mathcal{C} + \lambda_2\mathcal{C}$ is convex.

Exercise 4.9

Show that f from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$ is convex if and only if its epigraph

$$epi(f) = \{(x, t) \in dom(f) \times \mathbb{R} \text{ such that } t \geq f(x)\},$$

is convex.

CHAPTER 5

Convexity and non-differentiability

In this chapter, we introduce the tools usefull for non-smooth convex optimization, such as the sub-differential and the proximal operator.

♡ 5.1 Lower semi-continuity

In this section, we do not suppose that the function f is regular, we will see however that any convex function has to have some kind of regularity.

♡ 5.1.1 Definition of lower semi-continuity

The take home message of this section is that a convex function can be very wild **on the boundary of its domain**, and that it is regular (almost differentiable) on the **interior** of its domain. To understand how wild a convex function can be, take **ANY** function $\phi : \mathbb{R}^2 \to \mathbb{R}^+$ and build the fonction

$$f(x,y) = \begin{cases} 0 & \text{if } x^2+y^2 < 1 \\ \phi(x,y) & \text{if } x^2+y^2 = 1 \\ +\infty & \text{if } x^2+y^2 > 1 \end{cases} \tag{5.1}$$

Then the function f is convex. Since ϕ is arbitrary (also non-negative), the behavior of f is arbitrary complex. Hence we need a notion of regularity, the correct notion for optimization is called **lower semi-continuity**.

Definition 5.1.1 — Lower semi-continuous function. A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be lower semi-continuous (l.s.c.) iff :

$$\forall (x_k)_k \in \mathbb{R}^n, \text{ if } \lim_{k\to+\infty} x_k \text{ exists, then } \liminf_k f(x_k) \geq f(\lim_k x_k).$$

We recall that for any sequence $(u_k)_k$, then $\liminf u_k$ always exists as the limit of the increasing sequence $(v_k)_k$ where $v_k = \inf_{p\geq k} u_p$. Similarly, it holds that $\limsup$ always

exists, that $\limsup \geq \liminf$ and that the $\lim u_k$ exists if and only if the $\liminf$ and $\limsup$ are equal. An other definition of lower semi continuity is

Definition 5.1.2 — Lower semi-continuous function. A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be lower semi-continuous (l.s.c.) iff:

$$\forall (x_k)_k \in \mathbb{R}^n, \text{ if } \lim_{k\to+\infty} x_k \text{ and } \lim_{k\to+\infty} f(x_k) \text{ exist, then } \lim_k f(x_k) \geq f(\lim_k x_k).$$

In practice, we prefer the geometrical interpretation of lower semi-continuity

Proposition 5.1.1 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. The following propositions are equivalent:

i. f is l.s.c
ii. For each $\alpha \in \mathbb{R}$, the levelset $\{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ is closed.
iii. $epi(f)$ is closed.

Proof

To be done

Attention, every function which is continuous on its domain is not necessarly l.s.c. !! The domain must also be closed for the function to be l.s.c

Exercise 5.1

Under which conditions is the function f l.s.c ?

$$f(x) = \begin{cases} 0 & \text{si } x \in [0,1[\\ a \geq 0 & \text{si } x = 1 \end{cases}$$

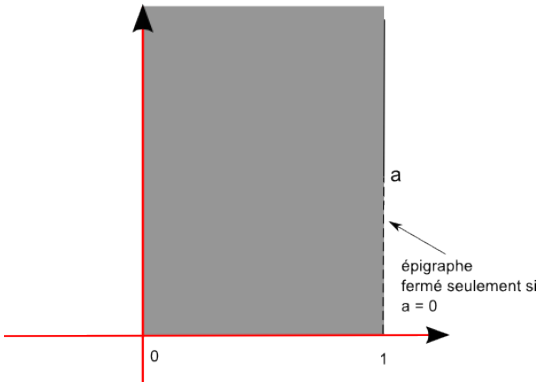


Figure 5.1: An example of function whose epigraph is open. Here ,the epigraph is closed if and only if $a = 0$.

Exercise 5.2

1. The convex functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that $dom(f) = \mathbb{R}^n$ are l.s.c.
2. The linear mappings are convex and l.s.c.

3. The functions $f(x) = \|x\|$ where $\|\cdot\|$ is any norm is convex and l.s.c.
4. The function (5.1) is always convex but l.s.c iff $\phi = 0$.

Exercise 5.3

Show that the ceil function (returns the closest integer greater than or equal to the input variable) is l.s.c. but not convex. Show that the floor function (returns the closest integer smaller than or equal to the input variable) is neither convex nor l.s.c.

♡ 5.1.2 Operations on l.s.c. functions

Theorem 5.1.2 Let f_1 and f_2 be two l.s.c functions and $\beta > 0$. The following functions f are l.s.c. :

1. $f(x) = \beta f_1(x)$.
2. $f(x) = (f_1 + f_2)(x)$ and $dom(f) = dom(f_1) \cap dom(f_2)$.
3. $f(x) = \max\{f_1(x), f_2(x)\}$ and $dom(f) = dom(f_1) \cap dom(f_2)$.

Proof

To be done (a little bit difficult...).

Theorem 5.1.3 Let ϕ be a convex l.s.c. function on $\mathbb{R}^m$, $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping and $b \in \mathbb{R}^n$. Then the function $f : x \mapsto \phi(Ax + b)$ is convex and l.s.c. and $dom(f) = \{x \in \mathbb{R}^n, Ax + b \in dom(\phi)\}$.

Proof

Let $x_1, x_2 \in dom(f)$. Denote $y_1 = Ax_1 + b$ and $y_2 = Ax_2 + b$. Then for every $\alpha \in [0, 1]$:

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= \phi(\alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b)) \\ &\leq \alpha\phi(Ax_1 + b) + (1-\alpha)\phi(Ax_2 + b) \\ &\leq \alpha f(x_1) + (1-\alpha)f(x_2). \end{aligned}$$

The epigraph is closed because the operator $x \mapsto Ax + b$ is continuous.

♡ 5.2 Continuity and differentiability of convex functions

This paragraph is devoted to studying the properties of continuity and differentiability of convex functions. A convex function may fail to be differentiable (think about the absolute value function $x \in \mathbb{R} \mapsto |x|$) but it has important regularity properties.

The first result is that every convex function is continuous on the interior of its domain

Theorem 5.2.1 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x_0 \in Int(dom(f))$. Then f is locally bounded, continuous and locally Lipschitz in x_0.

Proof

Let f be a convex function and $x_0 \in Int(dom(f))$.

- **Boundeness** : We first prove that f is locally bounded. Let $\varepsilon > 0$ such that for each $i = 1, \dots, n$, $x_0 \pm \epsilon e_i \in dom(f)$. Compute $M = \max_i f(x_0 \pm \varepsilon e_i)$. Consider the ball B centered at x_0, of radius ε for the ℓ_1 norm. Any $x \in B$ can be written as

$$x = x_0 + \sum_{i=1}^{n} \alpha_i e_i \quad \text{with} \quad \sum_{i=1}^{n} |\alpha_i| \leq \varepsilon$$

  Denote $\beta \in [0, 1[$ such that $\sum_{i=1}^{n} |\alpha_i| = \beta\varepsilon$ and write

$$x = \sum_{i=1}^{n} |\alpha_i| \left(x_0 + \varepsilon \frac{\alpha_i}{|\alpha_i|} e_i \right) + \frac{1-\beta}{2} \left(x_0 + \varepsilon \frac{\alpha_1}{|\alpha_1|} e_1 \right) + \frac{1-\beta}{2} \left(x_0 - \varepsilon \frac{\alpha_1}{|\alpha_1|} e_1 \right).$$

  Then x is written as a convex combination of elements of $(x_0 \pm \varepsilon e_i)_i$, and it holds that

$$f(x) \leq M$$

- **Lipschitz continuity** : Let $y \neq x_0$ such that $y \in B$. We denote $\alpha \in]0, 1[$ such that $\|y - x_0\|_1 = \alpha\varepsilon$. We denote $z = x_0 + \frac{1}{\alpha}(y - x_0)$. We have $\|z - x_0\| = \varepsilon$, so that $z \in B$. By construction, we have $0 \leq \alpha \leq 1$ and $y = \alpha z + (1 - \alpha)x_0$. The convexity of now implies:

$$\begin{aligned} f(y) &\leq (1-\alpha) f(x_0) + \alpha f(z) \\ &\leq f(x_0) + \alpha(M - f(x_0)) \\ &= f(x_0) + \frac{M - f(x_0)}{\varepsilon} \|y - x_0\|. \end{aligned}$$

  Hence there exists a constant C (independent of y) such that

$$f(y) - f(x_0) \leq C\|y - x_0\|.$$

  In order to obtain the second inequality, let $u = x_0 + \frac{1}{\alpha}(x_0 - y)$. We have $\|u - x_0\| = \epsilon$ so that $u \in B$ and $x_0 = \frac{1}{\alpha+1} y + \frac{\alpha}{\alpha+1} u$. We use the coercivity of f

$$\begin{aligned} \frac{1}{\alpha+1} f(y) + \frac{\alpha}{\alpha+1} f(u) &\geq f(x_0) \\ f(y) - f(x_0) &\geq -\alpha(f(u) - f(x_0)) \\ &\geq -\frac{M - f(x_0)}{\varepsilon} \|y - x_0\|. \end{aligned}$$

  We finally get that there exists C such that $\forall y \in B$:

$$|f(y) - f(x_0)| \leq C\|y - x_0\|.$$

An important consequence is that whatever prevents a convex function to be l.s.c. must be located at the boundary of the function.

We turn our attention to proving differentiability of a convex function. We first have to recall the notion of directional derivative

Definition 5.2.1 — Directional derivative. Let $x \in dom(f)$. We call **directional derivative** of f at point x in the direction d, the following limit (if it exists) :

$$f'(x;d) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}(f(x+\alpha d) - f(x)).$$

Whenever it exists, the directional derivative yields information on the slope of the function in the direction d, exactly like the 1D case.

- If $f'(x;d) > 0$, then f is increasing in the direction d.
- If $f'(x;d) = 0$, then no conclusions can be drawn.
- si $f'(x;d) < 0$, then f is decreasing in the direction d.

Attention ! Even if a function is directionally derivable in every direction, it may fail to be differentiable, even in 1D. The typical counter-example is the function $x \to |x|$. At the point 0, it admits a derivative on the right and on the left, but it is not derivable at 0. A remarkable theorem is that a convex function admits derivative in every direction.

Theorem 5.2.2 A convex function admits derivative in every direction for every point in the interior of its domain.

Proof

Let $x \in Int(dom(f))$ and $d \in \mathbb{R}^n$. Consider the function

$$\phi(\alpha) = \frac{1}{\alpha}(f(x+\alpha d) - f(x)), \alpha > 0.$$

There exists a ε such that $x + \epsilon d \in dom(f)$, so that the function ϕ is defined for every $\alpha \in]0, \varepsilon]$. Let $0 < \beta < 1$, we have :

$$f(x+\alpha\beta d) = f((1-\beta)x + \beta(x+\alpha d)) \leq (1-\beta)f(x) + \beta f(x+\alpha d).$$

And

$$\phi(\alpha\beta) = \frac{1}{\alpha\beta}(f(x+\alpha\beta d) - f(x)) \leq \frac{1}{\alpha}(f(x+\alpha d) - f(x)) = \phi(\alpha).$$

The function ϕ is decreasing in a neighbourhood of 0^+. Because f is L-Lipschitz, then ϕ is bounded. Hence ϕ is decrasing and bounded from below and the limit of $\phi(\alpha)$ as $\alpha \to 0^+$ exists.

The main interest of the above result is that we are able to compute directions of descent of convex functions at any point in the interior of its domain. Even if the function is not differentiable.

Lemma 5.2.3 Let f be a convex function and $x \in Int(dom(f))$. Then:

1. The function $d \mapsto f'(x;d)$ is convex and homogenous of degree 1 i.e.:

$$\forall \alpha \in \mathbb{R},\ \forall d \in \mathbb{R}^n,\ f'(x;\alpha d) = \alpha f'(x;d).$$

2. For all $y \in dom(f)$ we have

$$f(y) \geq f(x) + f'(x; y-x). \tag{5.2}$$

Proof

We first prove homegeneity, for every $p \in \mathbb{R}^n$ and $\tau > 0$, we have

$$\begin{aligned} f'(x,\tau p) &= \lim_{\alpha \to 0^+} \frac{1}{\alpha}(f(x+\alpha\tau p) - f(x)) \\ &= \tau \lim_{\beta \to 0^+} \frac{1}{\beta}(f(x+\beta p) - f(x)), \text{ if we set } \beta = \tau\alpha \\ &= \tau f'(x_0, p). \end{aligned}$$

Moreover, for every p_1, $p_2 \in \mathbb{R}^n$ and $\beta \in [0,1]$, we have

$$\begin{aligned} f'(x, \beta p_1 + (1-\beta)p_2) &= \lim_{\alpha \to 0^+} \frac{1}{\alpha}\left[f(x + \alpha(\beta p_1 + (1-\beta)p_2)) - f(x)\right] \\ &\leq \lim_{\alpha \to 0^+} \frac{1}{\alpha}\left\{\beta[f(x+\alpha p_1) - f(x) + (1-\beta)[f(x+\alpha p_2) - f(x)]\right\} \\ &= \beta f'(x, p_1) + (1-\beta) f'(x, p_2). \end{aligned}$$

The function $f'(x,p)$ is convex in p. To finish, we proceed as in the proof of Theorem 4.2.1. We start from

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

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 (5.2).

♡ 5.3 Sub-differential of a convex function

♡ 5.3.1 Definition and properties

Definition 5.3.1 — Sub-gradient and sub-differential. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x_0 \in dom(f)$. The **sub-differential** of f in x_0 is the set denoted by $\partial f(x_0)$ and defined by:

$$\partial f(x_0) = \left\{g \in \mathbb{R}^n \mid \forall x \in dom(f),\ f(x) \geq f(x_0) + \langle g, x - x_0\rangle\right\}.$$

Every vector $g \in \partial f(x_0)$ is a called a **sub-gradient** of f at the point x_0.

The graphical interpretation of the sub-differential is as follows :

1. Draw the graph and the epigraph of the function in $\mathbb{R}^{n+1}$.
2. For any point $(x_0, f(x_0)) \in \mathbb{R}^{n+1}$ on the graph, chose an hyperplane whose intersection with the epigraph is reduced to $(x_0, f(x_0))$.
3. Take the normal vector of this hyperplane and suppose that its last coordinate is non-zero (the hyperplane is not vertical). Then there exists a normal vector of the form $(g, 1)$.
4. The vector $g \in \mathbb{R}^n$ is a sub-gradient. The sub-differential is exactly the set of sub-gradient which are obtained by this method.

Those hyperplanes are called the **supporting hyperplanes** ("hyperplans supports" in french). Considering wether vertical hyperplanes are supporting hyperplanes or not depends on the author.

Proposition 5.3.1 — An other definition of the sub-differential. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x_0 \in dom(f)$. Then:

$$\partial f(x_0) = \{g \in \mathbb{R}^n \mid \forall d \in \mathbb{R}^n,\ f'(x; d) \geq \langle g, d \rangle\}.$$

Proof

To be done.

Exercise 5.4

Compute the sub-differential of $f : x \in \mathbb{R} \mapsto |x|$ in any point $x \in \mathbb{R}$.

Solution of Exercise 5.4

We find

$$\partial f(x) = \begin{cases} \{-1\} & \text{if} \quad x < 0 \\ \{+1\} & \text{if} \quad x > 0 \\ [-1, 1] & \text{if} \quad x = 0 \end{cases}$$

Lemma 5.3.2 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x_0 \in Int(dom(f))$. Then:

1. The sub-differential $\partial f(x_0)$ is a closed convex set.
2. If f is also l.s.c, then $\partial f(x_0)$ is non-empty and bounded.

Proof

1. Let g_1 and g_2 be two sub-gradients of $\partial f(x_0)$. We have $\forall y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle g_1, y - x \rangle$$
$$f(y) \geq f(x) + \langle g_2, y - x \rangle.$$

For any $\alpha \in [0, 1]$, multiply the first inequality by α, the second by $(1 - \alpha)$, sum up the two inequalities, then $\alpha g_1 + (1 - \alpha) g_2 \in \partial f(x_0)$. In order to prove that the sub-differential is closed, take a sequence g_n of

sub-gradients that converges to some g. For each y and n it is true that

$$f(y) \geq f(x) + \langle g_n, y - x \rangle.$$

Let n goes to $+\infty$, this proves that $g \in \partial f(x_0)$.

2. (Higly non-trivial).

Remark 5.3.3 The sub-differential may fail to exist on the boundary of the domain. For instance the function $f(x) = -\sqrt{x}$ on $\mathbb{R}_+$ is closed and convex, but the sub-differential does not exists on 0 because $\lim_{x \to 0^+} f(x) = -\infty$.

♡ 5.3.2 Properties of the sub-differential

Lemma 5.3.4 Let f be a convex l.s.c. function. Then

i. If f is differentiable at $x_0 \in Int(dom(f))$, then :

$$\partial f(x_0) = \{\nabla f(x_0)\}.$$

ii. If $dom(f) \subset \mathbb{R}^n$, $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator and $b \in \mathbb{R}^n$, then the function $\phi(x) = f(Ax + b)$ est convex l.s.c and:

$$\forall x \in Int(dom(\phi)),\ \partial\phi(x) = A^\top \partial f(Ax + b).$$

iii. If $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ with f_1 and f_2 convex and l.s.c. on $\mathbb{R}^n$ and $(\alpha_1, \alpha_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Then:

$$\begin{aligned} \partial f(x) &= \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) \\ &= \{\alpha_1 g_1 + \alpha_2 g_2 \mid (g_1, g_2) \in \partial f_1(x) \times \partial f_2(x)\}. \end{aligned}$$

iv. Let $g : \mathbb{R} \cup \{+\infty\} \to \mathbb{R} \cup \{+\infty\}$ be a convex increasing function, let $h = g \circ f$. Then:

$$\forall x \in int(dom(f)),\ \partial h(x) = \{\eta_1 \eta_2 \mid \eta_1 \in \partial g(f(x)), \eta_2 \in \partial f(x)\}. \tag{5.3}$$

Exercise 5.5

1. Let $f(x, y) = x + 2|y|$. Show that f is convex and l.s.c. and compute its sub-differential.
2. Let $f(x) = \|x\|_2$. Show that f is convex and l.s.c., compute $\partial f(x)$ for all $x \in \mathbb{R}^n$.
3. Let $g(t) = \frac{1}{2}t^2$ and $h(x) = g(f(x)) = \frac{1}{2}\|x\|_2^2$. Show that:

$$\partial f(x) = \begin{cases} \left\{\frac{x}{\|x\|_2}\right\} & \text{si } x \neq 0, \\ \{x, \|x\|_2 \leq 1\} & \text{sinon.} \end{cases} \tag{5.4}$$

Lemma 5.3.5 Let $(f_i)_{i=1..m}$ be a set of l.s.c. functions. Then the function

$$f(x) = \max_{i=1..m} f_i(x)$$

is also convex and l.s.c., its domain is $dom(f) = \cap_{i=1}^m dom(f_i)$ and:

$$\forall x \in Int(dom(f)),\ \partial f(x) = \text{conv}\left((\partial f_i(x), i \in I(x))\right)$$

where $I(x) = \{i \in \{1, .., m\}, f_i(x) = f(x)\}$.

♡ 5.3.3 Exemple 1: Sub-differential of an indicatrix function

Definition 5.3.2 Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex closed set. Let $x_0 \in \mathcal{C}$. We define the **normal cone** of $\mathcal{C}$ at x_0 the set $N_{\mathcal{C}}(x_0)$ given by :

$$N_{\mathcal{C}}(x) = \{\eta \in \mathbb{R}^n, \langle \eta, y - x\rangle \leq 0, \forall y \in \mathcal{C}\}.$$

Two examples of normal cones are given in Figure 5.3.3.

Figure 5.2: Exemples of normal cone. On the left : normal cone on a regular point of the boundary. On the right : normal cone on a singularity.

Lemma 5.3.6 — Sub-differential of an indicatrix function. Let $\mathcal{C} \in \mathbb{R}^n$ be a closed convex set. The indicatrix of $\mathcal{C}$ is defined as :

$$\mathbb{1}_X(x) = \begin{cases} 0 & \text{si } \ x \in X \\ +\infty & \text{sinon} \end{cases}$$

The sub-differential of $\mathbb{1}_{\mathcal{C}}$ at a point $x \in \partial\mathcal{C}$ is exactly the normal cone:

$$\forall x \in \mathcal{C}, \partial \mathbb{1}_{\mathcal{C}}(x) = N_{\mathcal{C}}(x).$$

Proof

♡ 5.3.4 Exemple 2: Sub-differential of a norm

Definition 5.3.3 — Dual norm. Let $\|\cdot\|_X$ be a norm on $\mathbb{R}^n$. We define the **dual norm** and we denote by $\|\cdot\|_{X^*}$, the norm defined by :

$$\|y\|_{X^*} = \sup_{x \in \mathbb{R}^n, \|x\|_X \leq 1} \langle x, y\rangle. \tag{5.5}$$

By definition of the dual norm, we have for every $(x, y) \in \mathbb{R}^{n\times n}$:

$$|\langle x, y\rangle| \leq \|x\|_X \|y\|_{X^*}, \tag{5.6}$$

which are generalized Cauchy-Schartz and Hölder's inequalities.

Proof

We prove that $\|\cdot\|_{X^*}$ is indeed a norm on $\mathbb{R}^n$.

Exercise 5.6

Let $\|x\|_X = \|x\|_p$ where $\|x\|_p$ is the usual ℓ^p norm on $\mathbb{R}^n$. Show that

$$\|y\|_{X^*} = \|x\|_{p^\star},$$

with $1/p + 1/p^\star = 1$. In particular, the ℓ^2 is its own dual norm.

Definition 5.3.4 — Sub-differential of a norm. Let $f(x) = \|x\|_X$, where $\|\cdot\|_X$ is a norm on $\mathbb{R}^n$. Then :

$$\partial f(x) = \underset{\|y\|_{X^*}\leq 1}{argmax}\langle x, y\rangle. \tag{5.7}$$

Proof

♡ 5.4 Necessary and sufficient condition of optimality

The notion of sub-differential allows us to caracterize the set of minima of a function

$$\text{Find} \quad x^* \text{ such that } f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$$

where f is a convex l.s.c. function

Theorem 5.4.1 Let $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a convex function. The point $x^* \in \mathbb{R}^n$ is a global minimum of f if and only if:

$$0 \in \partial f(x^*).$$

Proof

If $0 \in \partial f(x^*)$, then $f(x) \geq f(x^*) + \langle 0, x - x^*\rangle = f(x^*),\ \forall x \in dom(f)$. Reciprocally if $f(x) \geq f(x^*), \forall x \in dom(f)$, then $0 \in \partial f(^*)$ by definition of the sub-differential

From the above lemma, we can recover the KKT theorem which is an equivalence in the convex setting.

Theorem 5.4.2 Let $(f_i)_{0\leq i\leq m}$ be a family of convex differentiable family. And suppose that Slater's conditions holds, that is there exists a point $\bar{x} \in \mathbb{R}^n$ such that $f_i(\bar{x}) < 0$ for every i. Then x^* is a global minimizer of :

$$\min_{x\in X} f_0(x), \quad X = \{x \in \mathbb{R}^n \text{ such that } f_i(x) \leq 0,\ 1 \leq i \leq m\} \tag{5.8}$$

iff then there exists $\lambda_i \geq 0$ such that :

$$\nabla f_0(x^*) + \sum_{i \in I^*} \lambda_i \nabla f_i(x^*) = 0, \tag{5.9}$$

where $I^* = \{i \in \{1, \dots, m\},\ f_i(x^*) = 0\}$.

Proof

The trick is to rewrite the problem (5.8) into :

$$\min_{x \in \mathbb{R}^n} \phi(x) \quad \text{où} \quad \phi(x) = \max(f_0(x) - f^*, f_1(x), \dots, f_m(x)), \tag{5.10}$$

with $f^* = f(x^*)$. It is easy to check that the above problem is equivalent to (5.8) when we prove that ϕ verifies the following properties

- $\phi(x) \geq 0$ for any $x \in X$ because x^* is the minimizer of f_0 on X. Moreover $\phi(x) = 0$ on X iff x is a minimizer of f_0.
- $\phi(x) > 0$ if $x \notin X$ because at least one of the f_i verifies $f_i(x) > 0$.

The function ϕ is convex as the maximum of convex functions and the conditions of optimality of (5.8) are $x^* \in \partial\phi(x)$. According to Lemma (5.3.5), this is equivalent to the existence of positive numbers $(\bar{\lambda}_i)_{0 \leq i \leq m}$ such that :

$$\bar{\lambda}_0 \nabla f_0(x^*) + \sum_{i \in I^*} \bar{\lambda}_i \nabla f_i(x^\star) = 0 \quad \text{where} \quad \bar{\lambda}_0 + \sum_{i \in I^*} \bar{\lambda}_i = 1. \tag{5.11}$$

It remains to prove that $\bar{\lambda}_0 > 0$. Suppose it is the case, then :

$$\sum_{i \in I^*} \bar{\lambda}_i \nabla f_i(x^\star) = 0. \tag{5.12}$$

This means that $x^\star$ is a critical point of the convex function $g(x) = \sum_{i \in I^*} \lambda_i f_i(x)$ and hence a global minimizer. This cannot be the case because of the Slater condition which ensures that $g(\bar{x}) < 0$. Hence $\bar{\lambda}_0 > 0$ and it is sufficient to set $\lambda_i = \bar{\lambda}_i / \bar{\lambda}_0$ for every $i \in I^*$ to end the proof.

♡ 5.5 Fenchel transform

The convex conjugation, also called Fenchel transform or Legendre-Fenchel transform is used :

- to convexify a function (by computing the bi-conjugate, i.e. the conjugate of the conjugate).
- to compute the sub-differential of a convex function.
- to compute **dual** problems. These **dual** problems bring information to the original problems (called the **primal** problems).
- to switch from Lagrangian mechanics to Hamiltonian mechanics.

This transform was introduced by Mandelbrojt in 1939, made precise and improved by Fenchel en 1949. This transform generalises the Legendre transform (1787).

Definition 5.5.1 — Convexe conjugate. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any function. The convex conjugate of f is given by:

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \langle s, x \rangle - f(x). \tag{5.13}$$

The mapping $f \mapsto f^*$ is called the Legendre-Fenchel transform. The function f^* is called the convex conjugate of f, Fenchel transform of f or Legendre-Fenchel transform of f.

The ideas behind the Fenchel transform is given in Figure 5.5.

Figure 5.3: $f^*(u)$ is the supremum of the vertical difference between the graphe of f and the one of the hyperplane defined by the linear map $\langle \cdot, u \rangle$.

The objective of the Legendre transform is to compute the set of affine minoration of f. Given a vector s, a affine minoration is a function $x \mapsto \langle s, x \rangle + \alpha$ such that

$$\langle s, x \rangle + \alpha \leq f(x). \quad \forall x$$

The obective is to determine, given s, the set of α that give an affine minoration. The largest possible α must verify

$$\alpha \leq f(x) - \langle s, x \rangle. \quad \forall x \Rightarrow \alpha \leq \inf_x f(x) - \langle s, x \rangle$$

In other words, for each $\alpha \leq -f^*(s)$, then $x \mapsto \langle s, x \rangle + \alpha$ is an affine minoration of f.

Proposition 5.5.1 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. then:

1. The function f^* is convex and l.s.c.
2. For every $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$:

$$f(x) + f^*(s) \geq \langle x, s \rangle.$$

Proof

The function f^* is convex, as a supremum of convex functions. It is l.s.c because its epigraph is the intersection of closed sets.

Theorem 5.5.2 — Bi-conjugate. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any function. Its biconjugate is the function defined by :

$$f^{**}(x) = \sup_{s \in \mathbb{R}^n} \langle x, s \rangle - f^*(s). \tag{5.14}$$

The biconjugate is the largest convex l.s.c. function which is smaller than f. In other words the epigraph of the biconjugate is the closed convex hull of the

epigraph of f.

Proof

Let $\Sigma \subset \mathbb{R}^n \times \mathbb{R}$ be the set of couples (s, α) so that the affine function $x \mapsto \langle s, x\rangle - \alpha$ is below f:

$$\begin{aligned}(s,\alpha) \in \Sigma &\Leftrightarrow f(x) \geq \langle s, x\rangle - \alpha,\ \forall x \in \mathbb{R} \\ &\Leftrightarrow \alpha \geq \sup_{x\in\mathbb{R}^n} \langle s, x\rangle - f(x) \\ &\Leftrightarrow \alpha \geq f^*(x),\ (\text{and } s \in dom(f^*)).\end{aligned}$$

We then have for every $x \in \mathbb{R}^n$,

$$\begin{aligned}\sup_{(s,\alpha)\in\Sigma} \langle s, x\rangle - \alpha &= \sup_{s\in dom(f^*),\ -\alpha\leq -f^*(s)} \langle s, x\rangle - \alpha \\ &= \sup_{s\in dom(f^*)} \langle s, x\rangle - f^*(s) \\ &= f^{**}(x).\end{aligned}$$

From a geometrical perspective, the epigraphs of affine funcitons associated to $(s, \alpha) \in \Sigma$ are the closed half-spaces containing $epi(f)$. The epigraph of their supremum is the closed convex hull of $epi(f)$.

Theorem 5.5.3 The biconjugate of f satisfies $f^{**} = f$ if and only if f is convex and l.s.c.

Proof

This is a direct consequence of 5.5.2.

Exercise 5.7: Some exemples of convex conjugates.

1. Let $p \in]1, +\infty[$ and q the conjugate exponent of p (that is such that $1/p + 1/q = 1$). Then:
$$(1/p|\cdot|^p)^* = 1/q|\cdot|^q.$$
2. Let $Q \in \mathbb{R}^{n\times n}$ be a SDP matrix and $f(x) = \frac{1}{2}\langle x, Qx\rangle$, then:
$$f^*(x) = \frac{1}{2}\langle Q^{-1}x, x\rangle.$$
3. Let L be a subspace of $\mathbb{R}^n$. We consider the indicatrix of L :
$$f(x) = \mathbb{1}_L(x) = \begin{cases} 0 & \text{if } x \in L \\ +\infty & \text{otherwise}\end{cases} \tag{5.15}$$
Its convex conjugate is $f^* = \chi_{L^\perp}$.
4. Let $f(x) = \|x\|$ where $\|\cdot\|$ is a norm. Then :
$$f^*(s) = \chi_B(s) \tag{5.16}$$
where $B = \{s \in \mathbb{R}^n, \|s\|^* \leq 1\}$ and $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$.

Proposition 5.5.4 Let f be a convex l.s.c. function. For every $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$:

$$s \in \partial f(x) \Leftrightarrow x \in \partial f^*(s). \tag{5.17}$$

Proof

We have :

$$\begin{aligned} s \in \partial f(x) &\Leftrightarrow f^*(s) + f(x) = \langle s, x \rangle \\ &\Leftrightarrow f^*(s) + f^{**}(x) = \langle s, x \rangle \\ &\Leftrightarrow x \in \partial f^*(s). \end{aligned}$$

♡ 5.6 Exercise

III Theory of optimization

CHAPTER 6

Existence

The first question to tackle before solving a problem is to show that the problem has at least solution. Our first goal in this chapter is to give a framework in which we can ensure existence of minimum. Before showing that there exists a solution, we even have something else to do. Indeed, the very first thing to do before solving a problem is to understand it fully, and hence to define precisely what is a minimum.

♣ 6.1 Definitions of the minimum and the infimum

♣ 6.1.1 Infimum and minimum of a subset of $\mathbb{R}$

In order to define the minimum of a function, it is necessary to define the minimum value of a subset of $\mathbb{R}$. It is defined as

Definition 6.1.1 — Minimum. Let $A \subset \mathbb{R}$, we say that A admits a **minimum** (or minimum value) if there exists $m \in \mathbb{R}$ such that

$$m \in A \text{ and } \forall x \in A \text{ then } x \geq m.$$

In this case m is called the minimum of A, and is denoted $\min(A)$. A set can have at most one minimum

Exercise 6.1: Minimum

Find the minimum (if they exists) of

$$[0,1] \quad \text{and} \quad \{3\} \cup [5,7] \quad \text{and} \quad]0,1]$$

Solution of Exercise 6.1

- $\min([0,1]) = 0$
- $\min(\{3\} \cup [5,7]) = 3$

- $\min(]0,1])$ does not exists

The huge problem with the minimum is that we are not sure it exists. And in mathematics, we cannot talk about things that do not exist. "Wovon man nicht sprechen kann, darüber muß man schweigen. " (Ludwig Wittgenstein, Tractatus logico-philosophicus). That is why we introduce the notion of infimum which always exists. This notion requires the definition of a lower bound.

Definition 6.1.2 — Lower bound. Let $A \subset \mathbb{R}$, we say that m is a **lower bound** of A if and only if

$$\forall x \in A \text{ we have } x \geq m$$

The french translation of *lower bound* is **minorant**. With obvious change, we can define the **upper bound**, or **majorant** in french.

Exercise 6.2: Lower bound

1. Give the set of lower bounds of $[0,1]$.
2. Give the set of lower bounds of $]0,1]$.

Solution of Exercise 6.2

1. The set of lower bounds of $[0,1]$ is $]-\infty,0]$. Indeed, if m is a lower-bound then $m \leq 0$. On the other hand if $m \leq 0$, then m is a lower bound.
2. The set of lower bounds of $]0,1]$ is $]-\infty,0]$. See above.

As we see in the above exercise, the set of lower bound of A is the same as the one of its interior or of its closure. Moreover, it is easy to show that the set of lower bounds is always a closed set (a set that contains its boundary).

Definition 6.1.3 — Infimum. Let $A \subset \mathbb{R}$ be non empty, we say that m is the **infimum** of A if and only if m is the largest lower bound. In the case where the set of lower bounds of A is empty (in other words, A does not have lower bounds), we set the infimum of A to be $-\infty$. We have

$$m = \inf(A) \iff m = \max\{z, \text{such that } z \leq t \quad \forall t \in A\}$$

Exercise 6.3: Infimum

Give the infimum of the following sets

$$]0,1] \quad \text{and} \quad [5,7] \quad \text{and} \quad \{-n^2, n \in \mathbb{N}\} \quad \text{and} \quad \mathbb{R}^+ \setminus \mathbb{Q}$$

Solution of Exercise 6.3

- $\inf]0,1] = 0$. Indeed forall $x \in]0,1]$, then $x \geq 0$, so that 0 is a lower-bound. Suppose that there exists a lower-bound m with $m > 0$, then there exists $u \leq 1$ such that $0 < u < m$. So that m cannot be a lower-bound. Hence 0 is the largest lower-bound.

- $\inf[5,7] = 5$. The proof is almost the same than above.
- $\inf\{-n^2, n \in \mathbb{N}\} = -\infty$. Indeed suppose that there exists a lower-bound m, we then have $m \leq -n^2$ for each $n \in \mathbb{N}$, which is absurd.
- $\inf \mathbb{R}^+ \setminus \mathbb{Q} = 0$. The proof is almost the same than the first item.

There is a loophole in Definition 6.1.3, we need to show that the L, the set of lower bounds of A admits a maximum to define the infimum (recall that some sets do not admit maximum). The existence of a maximum is ensured by the fact that the set of lower bounds is closed and that A is non-empty. Indeed, any element of A is an upper-bound of L and in $\mathbb{R}$, every closed which has at least an upper-bound has a maximum. Do you know the proof of the last assertion? I am pretty sure you do not. Indeed this property is the *upper-bound axiom*, and is not to be proven (because it is an axiom).

The link between infimum and minimum that has to be bore in mind is the following

Proposition 6.1.1 — Infimum et minimum. Let A be a subset of $\mathbb{R}$.

1. The infimum of A always exists, the minimum of A may not exist.
2. If the minimum exists of A it is equal to the infimum of A .
3. If the infimum of A is equal to $-\infty$, there is no minimum of A .
4. The minimum of A exists if and only if the infimum of A belongs to A.

♣ 6.1.2 Minimum and infimum of a fonction

The notion of infimum and minimum of a subset of $\mathbb{R}$ allows us to define the infimum and minimum of a fonction from a set X to $\mathbb{R}$.

Definition 6.1.4 — Minimum and infimum. Let X be any set and $f : X \to \mathbb{R}$, we recall that $f(X)$ is a subset of $\mathbb{R}$ defined by:

$$f(X) = \{f(x), \text{ such that } x \in X\}$$

We define the infimum value of f over X, which we denote $\inf_{x\in X} f(x)$, as

$$\inf_{x\in X} f(x) = \inf f(X).$$

Moreover if the minimum of $f(X)$ exists, then there exists $x^\star \in X$ such that

$$f(x^\star) = \min f(X) = \inf f(X) = \inf_{x\in X} f(x)$$

Then we say that

- The infimum value of f over X is attained.
- We call this infimum value the **minimum value**.

- We call $x^\star$ a **minimizer** of f over X. They might be several $x^\star$. The set of minimizers is denoted arg min, hence

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

Proof

We show that if the minimum of $f(X)$ exists, there exists an $x^\star$ that verifies

$$f(x^\star) = \inf f(X).$$

Suppose indeed that the minimum of $f(X)$ exists, there exists $y \in f(X)$ such that $y = \inf f(X)$. Because $y \in f(X)$, then y is the image by f of some $x^\star$, that is $y = f(x^\star)$. Hence there exists $x^\star$ that verifies

$$f(x^\star) = \inf f(X).$$

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.

Exercise 6.4

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

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

Solution of Exercise 6.4

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. $\arg\min_{x\in X} f(x) = \emptyset$

♣ 6.1.3 Local minimum

A local minimizer is a point that minimizes a function compared to its nearest neighbours. The value of the function at a local minimizer is called a local minimum. The definition is as follows

Definition 6.1.5 — 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".

Exercise 6.5

If we set $X = [0, 1]$ and if the graph of f is given in Figure **??**, then local minimizers are plotted with orange circles. There are two global minimizers.

♠ 6.1.4 Mimizing sequences

Let $A \subset \mathbb{R}$, we say that $(x_n)_n$ is a minimizing sequence for A if $x_n \in A$ for every n and $\lim_{n\to+\infty} x_n = \inf(A)$. Let $f : X \mapsto \mathbb{R}$, 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 $\lim_{n\to+\infty} f(x_n) = \inf_X f$.

Proposition 6.1.2 If the value of the infimum is not $+\infty$, there always exists a minimizing sequence

Proof

- First consider $A \subset \mathbb{R}$, we search for a minimizing sequence for A. Suppose first that $\inf(A) = -\infty$, then for every $n \in \mathbb{N}, \exists x_n \in A$ such that $x_n < -n$, and then $(x_n)_n$ is a minimizing sequence. If $\inf(A) = m \in \mathbb{R}$, for every $n \in \mathbb{N}, \exists x_n \in A$ such that $x_n < m + \frac{1}{n}$ (because $m + \frac{1}{n}$ is **not** a lower bound of A). Since we always have $x_n \geq m$ (because m is a lower bound), then $(x_n)_n$ converges to m and $(x_n)_n$ is a minimizing sequence.
- Let X and f, we now search a minimizing sequence of $\inf_X f$. Consider $(y_n)_n$ a minimizing sequence of $f(X)$, then $y_n \to \inf_X f$ and for each $n, \exists x_n \in X$ such that $y_n = f(x_n)$. And we are done.

♣ 6.2 Existence of minimum in finite dimension

In this section, we give sufficient conditions that will 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 caracteristic of three different phenomenon that the hypotheses will prevent from occuring.

Proposition 6.2.1 The three following problems do not admit a global minimum :

- No continuity

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

- **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 6.2.2 — Existence of minimum. The function f admits a minimum on $X \subset \mathbb{R}^d$ if
- **continuity:** The function f is continuous on X.
- **closedness:** The set X is non-empty and closed (it contains its boundaries).
- **coercivity:** For every sequence $(x_n)_n \in X$ such that
$$\text{If } \lim_{n\mapsto+\infty} \|x_n\| = +\infty \quad \text{then} \quad \lim_{n\mapsto+\infty} f(x_n) = +\infty$$

Proof

1. By Proposition 6.1.2, because X is non-empty, there exists a 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
4. X is **closed** so that $x^\star \in X$.
5. f is **continuous** so that
$$\inf_X f = \lim_{n\mapsto+\infty} f(x_n) \underbrace{=}_{\text{continuity}} f(\lim_{n\mapsto+\infty} x_n) = f(x^\star)$$
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 it, 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.

Theorem 6.2.3 — Closedness of X when in standard form. Suppose that there exists a finite number of functions $g_i : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_j : \mathbb{R}^n \mapsto \mathbb{R}$ such that

$$X = \{x \text{ such that } g_i(x) \leq 0, h_j(x) = 0 \text{ for every } i = 1..I \quad j = 1..J\}$$

If every function g_i and h_j are continuous, then X is closed.

Proof

There are two ways to prove this, the first way is to state that if g_i and h_j are continous, then $g_i^{-1}(\mathbb{R}^-)$ and $h_j^{-1}(\{0\})$ are closed for every i and j. Hence

$$X = \cap_i g_i^{-1}(\mathbb{R}^-) \bigcap \cap_j h_j^{-1}(\{0\}),$$

is closed as the finite intersection of closed sets. The other way to prove Theorem 6.2.3 is to take a sequence $(x_n)_n$ of elements of X that converges to, say, $x^\star$. By continuty of g_i and h_j, we have $g_i(x^\star) \leq 0$ and $h_j(x^\star) = 0$ for every i and j. Hence X is closed.

We turn ourselves to proving coercivity assumption. There is usually two lines of proof to prove this assumption.

Theorem 6.2.4 — Ensuring coercivity. 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

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

Exercise 6.6

Show that the following problems admit a solution

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

Solution of Exercise 6.6

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 toprove coercivity, denoting $M = (x, y)$, we

have

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

and then

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

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

Exercise 6.7

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

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

Solution of Exercise 6.7

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 $n \mapsto +\infty$.

Remark 6.2.5 — How to improve the Theorem 6.2.2 ?.

- We can replace the coercivity hypothesis by : $\exists x_0 \in X$ such that for each sequence $(x_n)_n \in X$ with $\lim_{n\to+\infty} \|x_n\| = +\infty$ then

$$\limsup_{n\to+\infty} f(x_n) > f(x_0)$$

- We can replace the continuity hypothesis by : For every sequence $(x_n)_n$ that converges to some $x^\star$ such that $\lim f(x_n)$ exists, then

$$f(x^\star) \leq \lim_{n\to+\infty} f(x_n)$$

This property is called *lower semi-continuity*. For instance the following problem admits a minimizer $X = [-1,1]$ and $f(x) = x^2$ if $x \neq 0$ and $f(0) = -1$

♠ 6.3 Existence of minimum in the infinite dimension case

In infinite dimension, the bounded closed sets are not necessarily compact. The correct theorem would be

Theorem 6.3.1 — Existence of minimima. There exists a minimum to f over X if
- X is non-empty and compact.
- f is continuous.

Proof. Follow exactly the steps of the proof of Theorem 6.2.2, but this time, instead of proving that $(x_n)_n$ is bounded and hence admits a convergent subsequence, use directly that $(x_n)_n$ is a sequence of elements of X. Because X is compact, $(x_n)_n$ admits a convergent subsequence. ■

♠ 6.3.1 Counterexamples

Exercise 6.8

For the following problems, show that the hypotheses of Theorem 6.2.2 are verified (except for the finite dimension) but that the problems do not admit a minimizer.

- $X = \ell^2(\mathbb{R})$ and $f(x) = (\|x\|^2 - 1)^2 + \sum_{i=0}^{+\infty} \frac{x_i^2}{i+1}$
- $X = \{x \in \ell^2(\mathbb{R}) \text{ such that } \|x\| = 1\}$ and $f(x) = \sum_{i=0}^{+\infty} \frac{x_i^2}{i+1}$

Solution of Exercise 6.8

In both cases, for each $x \in X$, we have $f(x) > 0$. It is sufficient to exhibit a sequence $(x^n)_n \in \ell^2(\mathbb{R})$ such that $f(x^n)$ converges to 0. In order to achieve this goal, use the following sequence defined for every n by

$$\begin{cases} x_i^n = 0 & \text{if } i \neq n \\ x_i^n = 1 \text{ if } i = n \end{cases}$$

♠ 6.3.2 Existence

In the convex case, there is a very powerfull theorem.

Theorem 6.3.2 If f and X are convex, there exists a minimum to f over X if
- X is non-empty and closed in an Hilbert space.
- f is continuous for the norm of the Hilbert space
- For every $(x_n)_n \in X$, then

$$\lim_{n \to +\infty} \|x_n\| = +\infty \Rightarrow \lim_{n \to +\infty} f(x_n) = +\infty$$

Proof

A sequence $(v_n)_n$ is said to weakly converge to some $v^\star$ if for every v, we have

$$\lim_{n \to +\infty} \langle v_n - v^\star, v \rangle = 0.$$

A standard example of sequence $(v_n)_n$ that weakly converges to 0 but does not strongly converges is given by $V = \ell^2(\mathbb{R})$ an $v_n = (0, \dots, 0, 1, 0, \dots)$ whith the 1 located at the n^{th} place.
Any minimizing sequence is bounded and hence weakly converges up to a subsequence (Banach-Alaoglu theorem). Then K is closed for weak convergence and f is lower semi-continuous for weak convergence.

♣ 6.4 The effect of convexity

♣ 6.4.1 Globalization of minima

Theorem 6.4.1 — 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 the critical point of f are global minima 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:

$$\forall y \in X \text{ such that } \|y - x_0\| < r \text{ then } f(y) \geq f(x_0). \tag{6.1}$$

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

$$f(x_1) < f(x_0). \tag{6.2}$$

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 (6.1).

2. If x is a critical point, then $\nabla f(x) = 0$. Because f is convex, we use Theorem 4.2.1 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.

♣ 6.4.2 Strict convexity

Theorem 6.4.2 — 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).

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.

♣ 6.5 Exercise

♣ 6.5.1 Some exercises

Exercise 6.9

Show that the following problem $\inf_{M\in X} f(M)$ admits minima

1. $f : (x, y) \mapsto x^2 + y^2$, $X = \mathbb{R}^2$
2. $f : (x, y) \mapsto 3x$, $X = \{(x, y) \text{ such that } x^2 + y^2 \leq 1\}$
3. $f : (x, y) \mapsto x^4 + y^6 + x^2y^2$, $X = \mathbb{R}^2$
4. $f : (x, y) \mapsto 4x^2 + 6y^2 - 5xy$, $X = \mathbb{R}^2$

♠ 6.5.2 More exercises

Exercise 6.10

Let $f : (x, y) \mapsto \frac{1}{100}x^2 + 10^{10}y^2 - 100xy$, show that the problem $\inf_{M\in\mathbb{R}^2} f(M)$ admits minima.

CHAPTER 7

Caracterization

The objective of this chapter is to study the so-called *Euler conditions*. There are two kind of Euler conditions, the *necessary* Euler conditions and the *sufficient* Euler conditions.

The necessary Euler conditions are conditions that has to be met for a point to be a local minimizer. The sufficient Euler conditions are conditions that will ensure a point is a local minimizer. Of course the necessary and sufficient Euler conditions are not the same or else they would have been called *equivalent* conditions. Moreover, if the sufficient condition is met, then the necessary condition is valid. In other words it is easier to meet the necessary condition than the sufficient one.

♣ 7.1 Euler conditions in dimension 1

We suppose that the student has already studied the case of a function f from $\mathbb{R}$ to $\mathbb{R}$. The following proposition is just a refresher.

Proposition 7.1.1 — Euler conditions in the case $n = 1$. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a C^2 function, then

- If $x^\star$ is a local minimum of f on $\mathbb{R}$ then $f'(x^\star) = 0$ and $f''(x^\star) \geq 0$.
- If $x^\star$ is such that $f'(x^\star) = 0$ and $f''(x^\star) > 0$ then $x^\star$ is a local minimum of f.

This proposition is stated later in a more general setting, we give the proof at this time. This proposition leads to a search tactic of local minima in dimension 1.

Definition 7.1.1 — Searching for local extrema if $f : \mathbb{R} \mapsto \mathbb{R}$.

1. Ensure that f is a C^2 function.
2. Solve the equation $f'(x) = 0$ (critical point equations)
3. For any point x such that $f'(x) = 0$.

- If $f''(x) > 0$, then x is a local minimum.
- If $f''(x) < 0$, then x is a local maximum.
- If $f''(x) = 0$, we cannot conclude.

♣ 7.2 Euler conditions in finite dimension

We state and prove in this section the Euler conditions for real valued function from $\mathbb{R}^n$. The Euler conditions given in the theorem below are a generalization of Proposition 7.1.1

Theorem 7.2.1 Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a C^2 function, then

1. If $x^\star$ is a local minimum of f on $\mathbb{R}^n$ then $\nabla f(x^\star) = 0$ and $H[f](x^\star) \succeq 0$.
2. If $x^\star$ 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 $\mathbb{R}^n$.

Proof

Both propositions rely on the following second order Taylor expansion:

$$f(x^\star + h) = f(x^\star) + \langle \nabla f(x^\star), h\rangle + \frac{1}{2}\langle H[f](x^\star)h, h\rangle + o\big(\|h\|^2\big) \tag{7.1}$$

- First suppose that $x^\star$ is a local minimum of f over $\mathbb{R}^n$, then for every direction $d \in \mathbb{R}^n$ with $\|h\| = 1$ and every step $\varepsilon > 0$ small enough, we have $f(x^\star + \varepsilon d) \geq f(x^\star)$ and then (7.1) yields :

$$\varepsilon\langle \nabla f(x^\star), d\rangle + \frac{\varepsilon^2}{2}\langle H[f](x^\star)d, d\rangle + o\big(\varepsilon^2\big) \geq 0 \tag{7.2}$$

  Divide (7.2) by ε and let ε go to 0. We obtain $\langle \nabla f(x^\star), d\rangle \geq 0$ for each d. If we choose $d = -\nabla f(x^\star)$, we obtain $\|\nabla f(x^\star)\|^2 \leq 0$ and therefore $\nabla f(x^\star) = 0$ which is the first thing to prove. Then (7.2) becomes

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

  Divide now the above equation by ε^2 and let ε go to 0, then we have $H[f](x^\star) \geq 0$ by Proposition 2.4.1.
- Suppose now that $\nabla f(x^\star) = 0$ and $H[f](x^\star) > 0$. In this case (7.1) becomes

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

  By Proposition 2.4.1, we have

$$\langle H[f](x^\star)d, d\rangle \geq c\|d\|^2,$$

  and then

$$f(x^\star + h) - f(x^\star) \geq \varepsilon^2\left(\frac{c}{2} + o(1)\right)$$

By definition of $o(1)$, the right hand side of the above equation is positive for ε small enough. Hence for $h = \varepsilon d$ small enough, we have :

$$f(x^\star + h) \geq f(x^\star).$$

Hence $x^\star$ is a local minimum of f.

Theorem 7.2.1 allows defining a strategy for finding local extrema. It is the counterpart of the $1d$ case.

Definition 7.2.1 — Searching for local extrema if $f : \mathbb{R}^n \mapsto \mathbb{R}$.

1. Ensure that f is a C^2 function
2. Solve the system $\nabla f(x) = 0$ (critical point equation).
3. For every x that verifies $\nabla f(x) = 0$, compute the sign of the eigenvalues of $H[f](x)$.
   - If every eigenvalue of $H[f](x)$ is > 0, then x is a local minimum of f.
   - If every eigenvalue of $H[f](x)$ is < 0, then x is a local maximum of f.
   - If some eigenvalues of $H[f](x)$ are > 0 and some are < 0, then x is neither a local minimum nor a local maximum, it is called a **saddle-point**
   - If all the eigenvalues of $H[f](x)$ have the same sign but some are equal to zero, then we cannot conclude.

Proposition 7.2.2 — Trick of the trade. If A is a matrix, its determinant is the product of the eigenvalues whereas its trace is the sum of the eigenvalues. In dimension 2, the product and the sign of two numbers are enough information to determine easily the sign of the numbers. Hence in dimension 2, there is virtually no computation necessary in order to determine the sign of the eigenvalues.

Exercise 7.1

Find the local extrema of f over $\mathbb{R}^2$, if

1. $f(x, y) = x^2 + y^2$
2. $f(x, y) = -x^2 - 5y^2$
3. $f(x, y) = x^2 - y^2$
4. $f(x, y) = x^4 + y^2$
5. $f(x, y) = x^3 + y^2$

Solution of Exercise 7.1

We let $M = (x, y)$

1. Here $\nabla f(M) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$ and $H[f](M) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. The only point for which $\nabla f(M) = 0$ is $M = 0$ and $H[f](0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. The Hessian is definite positive. There is only one local extrema, located at $M = 0$, it is a local minimum.

2. Here $\nabla f(M) = \begin{pmatrix} -2x \\ -10y \end{pmatrix}$ and $H[f](M) = \begin{pmatrix} -2 & 0 \\ 0 & -10 \end{pmatrix}$. The only point for which $\nabla f(M) = 0$ is $M = 0$ and $H[f](0) = \begin{pmatrix} -2 & 0 \\ 0 & -10 \end{pmatrix}$. The Hessian is definite negative. There is only one local extrema, located at $M = 0$, it is a local maximum.
3. Here $\nabla f(M) = \begin{pmatrix} 2x \\ -2y \end{pmatrix}$ and $H[f](M) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$. The only point for which $\nabla f(M) = 0$ is $M = 0$ and $H[f](0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$. The Hessian has one positive and one negative eigenvalue. There is no local extrema, the point $M = 0$is a saddle point.
4. Here $\nabla f(M) = \begin{pmatrix} 4x^3 \\ 2y \end{pmatrix}$ and $H[f](M) = \begin{pmatrix} 12x^2 & 0 \\ 0 & 2 \end{pmatrix}$. The only point for which $\nabla f(M) = 0$ is $M = 0$ and $H[f](0) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$. The Hessian has one positive eigenvalue and one equal to zero. If there exists a local extrema, it is at the point $M = 0$ and it is a local minimum. We have to decide if $M = 0$ is a local minimum or not. We see that $f(0) = 0$ and that $f \geq 0$. Hence 0 is a global minimum of f (hence a local one).
5. Here $\nabla f(M) = \begin{pmatrix} 3x^2 \\ 2y \end{pmatrix}$ and $H[f](M) = \begin{pmatrix} 6x & 0 \\ 0 & 2 \end{pmatrix}$. The only point for which $\nabla f(M) = 0$ is $M = 0$ and $H[f](0) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$. The Hessian has one positive eigenvalue and one equal to zero. If there exists a local extrema, it is at the point $M = 0$ and it is a local minimum. We have to decide if $M = 0$ is a local minimum or not. We see that $f(0) = 0$ and that if x is small and negative, then $f(x, 0) < 0$. Hence 0 is not a local minimum of f.

♣ 7.3 A gentle introduction to the constrained case

In the constrained case, i.e., when f is not minimized of $\mathbb{R}^n$ but over $X \subset \mathbb{R}^n$, the Euler conditions are more complex. In a nutshell, the gradient of f might not cancel on a local extremum if this extremum is located on the boundary. The caracterization of the local extrema is coined as **the KKT theorem**. The goal of this section is to introduce this theorem in two simple cases We start by an example of the KKT theorem before stating the main result

Proposition 7.3.1 — Constraint minimization on $\mathbb{R}$.. Consider the problem over the constraint set $X = [a, b] \subset \mathbb{R}$. Suppose that $x^\star$ is a local minimum of f over

X, then the following alternatives must hold:

$$\begin{cases} f'(x^\star) \geq 0 \text{ if } x^\star = a \\ f'(x^\star) = 0 \text{ if } x^\star \in]a, b[\\ f'(x^\star) \leq 0 \text{ if } x^\star = b \end{cases}$$

In a nutshell, if a local minimum is in the interior of the domain, then its derivative must vanish, but if it is on the boundary of the domain, then we can only control the sign of the

Theorem 7.3.2 — Minimizing with one inequality constraint. Let f and g be C^1 functions from $\mathbb{R}^n$ to $\mathbb{R}$ and let $X = \{x, g(x) \leq 0\}$. Assume that for any point x such that $g(x) = 0$, we have $\nabla g(x) \neq 0$. Let $x^\star \in X$ be a local minimum of f over X. There exists $\lambda \in \mathbb{R}$ such that

$$\begin{cases} \nabla f(x^\star) + \lambda \nabla g(x^\star) = 0 \\ \lambda \geq 0 \text{ and } g(x^\star) \leq 0 \\ g(x^\star) = 0 \text{ or } \lambda = 0 \end{cases}$$

The conclusions of Theorem 7.3.2 can be interpreted as

- either $g(x^\star) < 0$ and then we must have $\nabla f(x^\star) = 0$
- or $g(x^\star) = 0$ and we must have

$$\nabla f(x^\star) + \lambda \nabla g(x^\star) = 0 \text{ with } \lambda \geq 0.$$

Proof

Let $x^\star$ be a local minimum of f over X. Let d with $\|d\| = 1$ be any direction $\mathbb{R}^n$. For any step $\varepsilon > 0$, perform a Taylor expansion of f and g around the point $x^\star$. It reads :

$$f(x^\star + \varepsilon d) = f(x^\star) + \varepsilon \langle \nabla f(x^\star), d \rangle + o(\varepsilon) \tag{7.3}$$
$$g(x^\star + \varepsilon d) = g(x^\star) + \varepsilon \langle \nabla g(x^\star), d \rangle + o(\varepsilon) \tag{7.4}$$

- If $g(x^\star) < 0$, then by continuity $g(x^\star + \varepsilon d) = g(x^\star) + o(1)$ so that

$$g(x^\star + \varepsilon d) < 0 \quad \forall \varepsilon \text{small enough.}$$

  Hence, for any d, $x^\star + \varepsilon d$ is in X for small enough ε. Since $x^\star$ is a local minimum of f over X, upon considering smaller ε, we must have $f(x^\star + \varepsilon d) \geq f(x^\star)$. And (7.3) yields:

$$\varepsilon \langle \nabla f(x^\star), d \rangle + o(\varepsilon) \geq 0$$

  As in the proof of the unconstrained case, we divide the above equation by ε, we then let ε go to 0 and we choose $d = -\nabla f(x^\star)$, we then obtain $\nabla f(x^\star) = 0$.
- Suppose now that $g(x^\star) = 0$ then we claim that it is impossible to find a direction d that verifies

$$\langle \nabla f(x^\star), d \rangle < 0 \text{ and } \langle \nabla g(x^\star), d \rangle < 0.$$

Indeed if such a direction existed, we would have, for small enough ε:

$$\begin{aligned} g(x^\star + \varepsilon d) &= \varepsilon\langle \nabla g(x^\star), d\rangle + o(\varepsilon) < 0 \\ f(x^\star + \varepsilon d) &= f(x^\star) + \varepsilon\langle \nabla f(x^\star), d\rangle + o(\varepsilon) < f(x^\star). \end{aligned}$$

And, for all small $\varepsilon > 0$, we could find points $x^\star + \varepsilon d \in X$ such that $f(x^\star + \varepsilon d) < f(x^\star)$. This contredicts the fact that $x^\star$ is a local minimizer of f over X. By assumption $\nabla g(x^\star) \neq 0$. If we suppose $\nabla f(x^\star) = 0$, there is nothing to prove in the theorem (take $\lambda = 0$). We restrict our attention to the case $\nabla f(x^\star) \neq 0$. We work in a 2-dimensional space that contains both vectors.In Figure 7.1 (left) we show in red the directions d such that $\langle \nabla g(x^\star), d\rangle < 0$ and in blue the directions d such that $\langle \nabla f(x^\star), d\rangle < 0$. The blue and red half-circles cannot intersect, so that $\nabla f(x^\star)$ and $\nabla g(x^\star)$ have to be aligned and in opposite directions, as in Figure 7.1 (right). It means that there exists $\lambda > 0$ tel que

$$\nabla f(x^\star) = -\lambda \nabla g(x^\star).$$

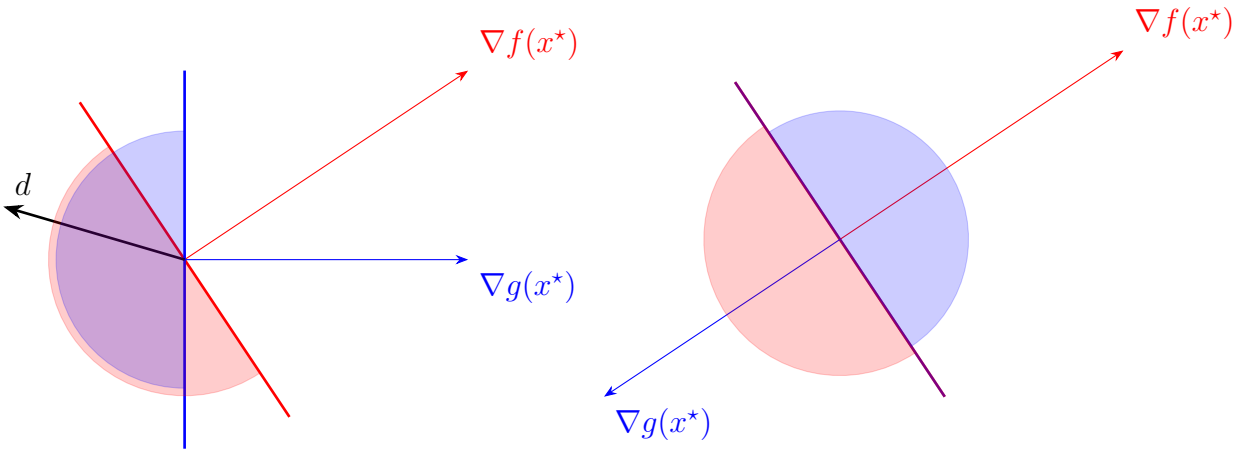


Figure 7.1: Proof of Theorem 7.3.2. On the left, we represent in black a direction d such that $\langle \nabla f(x^\star), d\rangle < 0$ et $\langle \nabla g(x^\star), d\rangle < 0$. This figure illustrates a simple application of Farkas lemma.

♠ 7.4 First order necessary conditions with constraints

In this chapter, we will deal with first order optimality conditions with constraints. We first state a very simple theorem, valid in the case of an open set of constraints.

Theorem 7.4.1 — X is open. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $X \subset \mathbb{R}^n$ be a open set. Suppose that $x^\star \in X$ is a local minimum of f on X, then $\nabla f(x^\star) = 0$.

The above theorem states that when X is open, the first order optimality conditions are the same than the one without constraints. Sadly, the case where X is open is of very small interest in practice. Indeed, in order to ensure existence of minimizers, the practitionner usually ensures that X is closed. In order to deal with closed sets, the most celebrated theorem is certainly the KKT theorem. The main problem with this theorem is that it relies on an hypothesis called **qualifications of constraints**. For pedagogical reasons, we leave this difficulty aside and discuss it further in Section **??**

Theorem 7.4.2 — Karush-Kuhn-Tucker (1951). Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^n \to \mathbb{R}^q$ be differentiable functions and let

$$X = \{x \in \mathbb{R}^n \text{ such that } h(x) = 0 \text{ and } g(x) \preceq 0\}.$$

Suppose that $x^\star \in X$ is a local minimum of f on X and suppose that **the constraints are qualified** at $x^\star$, then there exists $\lambda^\star \in \mathbb{R}^p$, and $\mu^\star \in \mathbb{R}^q$ such that :

$$\begin{cases} \nabla f(x^\star) + \sum_{i=1}^p \lambda_i^\star \nabla g_i(x^\star) + \sum_{j=1}^q \mu_j^\star \nabla h_j(x^\star) = 0 \\ h(x^\star) = 0 \text{ and } \lambda^\star \succeq 0, g(x^\star) \preceq 0, \lambda^\star \cdot g(x^\star) = 0 \end{cases}$$

Several things can be said about this theorem

- First, in the case where there is no constraints, the KKT theorem boils down to the standard theorem where $\nabla f(x^\star) = 0$. Hence, the KKT theorem is just an extension of the standard case.
- One may wonder what is the meaning of the sentence **the constraints are qualified**. As it happens, we need several hypothesis to make this theorem work. These hypothesis are linked to the fact that the problem has been posed in a correct way and they are just grouped under the name of qualification of constraints. One of our goal is to determine those hypothesis.
- The condition $\lambda^\star \cdot g(x^\star) = 0$ boils down to $\lambda_i = 0$ or $g_i(x^\star) = 0$ for all i.

We first study a simple example in order to clarify the KKT theorem.

Exercise 7.2

Find every rectangle in $\mathbb{R}^2$ of maximal surface with perimeter smaller than $P \in \mathbb{R}^{+*}$. You will assume that the constraints are qualified.

Solution of Exercise 7.2

Let a and b be the lengths of the rectangle if $x = (a, b)$ and denote

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

If $X = \{x \text{ such that } g(x) \preceq 0\}$ then we aim at minimizing f over X. The

function f is continuous and X is closed (by continuity of g). A drawing shows that X is a triangle, it is therefore bounded. There exists a minimum and a maximum of f on X. We suppose that the constraints are qualified, we write down KKT equations : Find x and λ such that $g_i(x) \leq 0$ and $\lambda_i \geq 0$ for $i \in \{1, 23\}$ such that :

$$\nabla f(x) + \sum_{i=1}^{3} \lambda_i \nabla g_i(x) = 0 \text{ and } \lambda_i g_i(x) = 0 \quad \forall i$$

The first equation is

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

- If $\lambda_1 = 0$, then $a = -\lambda_3$ and $b = -\lambda_2$. From $\lambda_3 g_3(x) = 0$ we deduce that either λ_3 or b is equal to 0, that is either a or b is zero. In any case $f = 0$.
- If $\lambda_1 \neq 0$ then we have $g_1(x) = 0$, hence $a + b = \frac{P}{2}$. If $\lambda_2 \neq 0$ (resp $\lambda_3 \neq 0$), then $0 = g_2(x) = -a$ (resp. $0 = b$) and $f = 0$. If $\lambda_2 = \lambda_3 = 0$ we have $a = b = \frac{\lambda_1}{2}$ and from $a + b = \frac{P}{2}$, we deduce that $a = \frac{P}{4}$ and $f = -\frac{P^2}{16}$.

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 $a = b$. The rectangle that maximises the area for a given perimeter is a square.

We now give without proof a sufficient condition to ensure qualification of constraints. At first we need to make the difference between **active** and **inactive** constraints. In a nutshell a constraint is inactive at a point x if it does not change locally (around x) the admissible set X.

Definition 7.4.1 — Active constraints. The inequality constraints $g_i(x) \leq 0$, $i \in \{1, \ldots, p\}$, is said to be **active** at x iff $g_i(x) = 0$. It is said to be **inactive** in x if $g_i(x) < 0$. For any $x \in X$, we denote

$$I(x) = \{i \in \{1, \ldots, n\} \text{ s.t. } g_i(x) = 0\},$$

the set of active constraints at x.

We can now state our main proposition about constraint qualification

Proposition 7.4.3 Let $X = \{x, g(x) \preceq 0, h(x) = 0\}$ and let $x^\star \in X$. If the family $\nabla h_j(x^\star)_{j=1,..,q} \cup (\nabla g_i(x^\star))_{i \in I(x^\star)}$ is linearly independent. Then the constraints are qualified at point $x^\star$.

Exercise 7.3

Prove that the constraints in the exercise 7.2 are qualified.

Solution of Exercise 7.3

We use Proposition 7.4.3. We recall that for any $x = (a, b) \in \mathbb{R}^2$, we have $g(x) = \begin{pmatrix} 2(a+b) - P \\ -a \\ -b \end{pmatrix}$. Let $x \in X$ we discuss according to the number of constraint that are active at x.

- No active constraint : there is nothing to do, the empty family is independent, hence the constraints are qualified at x.
- Exactly one active constraint : $\exists! i \in \{1, 2, 3\}$ such that $g_i(x) = 0$. We have

$$\nabla g_1(x) = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \text{ and } \nabla g_2(x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \text{ and } \nabla g_3(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

  We always have $\nabla g_i(x) \neq 0$, hence for any $i \in \{1, 2, 3\}$ the family $(\nabla g_i(x))$ is independent.
- Exactly two active constraint : In view of the previous calculations of the gradient, we pick any i and j in $\{1, 2, 3\}$ with $i \neq j$ and we manually check for the three different choices that $(\nabla g_i(x), \nabla g_j(x))$ is independent.
- Exactly two active constraint : We check that $g_1(x) = g_2(x) = g_3(x)$ is impossible.

We now showcase an example of application of the KKT Theorem

Exercise 7.4

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

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

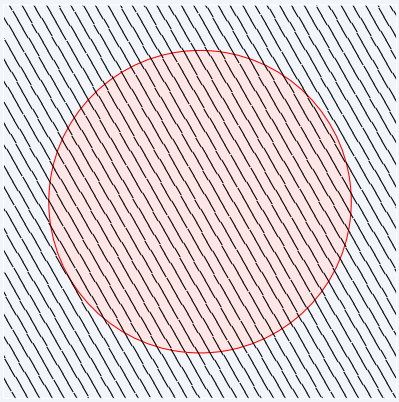
Solution of Exercise 7.4

We split the solution into several different simpler 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**



- **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** Let $M = (x, y) \in \mathbb{R}^2$, we discuss according to the number of constraints active at point M
  - No active constraints : nothing to do.
  - One active constraint, that is $g(M) = 0$, hence $x^2 + y^2 = 1$. We have $\nabla g(M) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$. For each M such that $g(M) = 1$, we have $\nabla g(M) \neq 0$, hence the family $(\nabla g(M))$ is independent.

  We proved that the constraints are qualified at each point.
- **Step 3 : Solving KKT** 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 global 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.

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

Proposition 7.4.4 — Slater's condition. Suppose that X is given by inequality conditions only and that the functions g_i are convex. If there exists a point $x \in X$ such that each constraints which is not affine is inactive at x, then the constraints are qualified everywhere in X.

Exercise 7.5

Show that the constraints are qualified in the set of constraints is $X = \{M \text{ s.t. } g(M) \leq 0\}$ if g is defined as

- $M = (a, b) \in \mathbb{R}^2$ and $g(M) = \begin{pmatrix} 2(a+b) - P \\ -a \\ -b \end{pmatrix}$ (see Exercise 7.2)
- $M = (x, y) \in \mathbb{R}^2$ and $g(M) = x^2 + y^2 - 1$ (see Exercise 7.4)

Solution of Exercise 7.5

- In this case, all the functions g_i are affine, the constraints are qualified everywhere.
- The function g is convex and $g(0) < 0$, hence the constraints are qualified everywhere.

Exercise 7.6

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

$$\min_{2x+y \geq 5} x^2 + y^2$$

Solution of Exercise 7.6

- **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

  $$f(M) = x^2 + y^2 \quad g(M) = -2x - y + 5.$$

  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**
- **Step 1 : Existence of a minimum** The function f is continuous and coercive and the set X is closed (because g is continuous).
- **Step 2 : Qualification of constraints** The function g is affine hence the constraints are qualified everywhere.

- **Step 3 : Solving KKT** KKT equations read

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

We discuss according to the cases in (3).

- **If $g(M) \neq 0$:** Then we must have $\lambda = 0$ by (3). Equation (1) yields $x = y = 0$ which is a contradiction with $g(M) \leq 0$.
- **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 $x = \frac{1}{\lambda}$ and $y = \frac{1}{2\lambda}$. From $g(M) = 0$, we have

$$\frac{2}{\lambda} + \frac{1}{2\lambda} = 5$$

And we have $\lambda = \frac{1}{2}$.We check that $\lambda \geq 0$ and we have a KKT point 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 global minimizer is a KKT point.
  3. There exists only one KKT point at point $\lambda = \frac{1}{2}$ and $x = 2$ and $y = 1$.
  4. There exists only one local minimizer at point $\lambda = \frac{1}{2}$ and $x = 2$ and $y = 1$. It is a global minimizer.

♠ 7.5 Second order optimality conditions with constraints

As in Section **??**, we start by stating the conditions in the case where the set X is open. This theorem is just an extension of Theorem 7.2.1 and the proof is similar.

Theorem 7.5.1 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.

There is a very usefull way to rewrite KKT theorem using the Lagrangian $\mathcal{L}$.

Theorem 7.5.2 — Karush-Kuhn-Tucker (1951). Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^n \to \mathbb{R}^q$ be C^2 functions and let

$$X = \{x \in \mathbb{R}^n \text{ such that } h(x) = 0 \text{ and } g(x) \preceq 0\}.$$

Define the Lagrangian as

$$\mathcal{L}(x,(\lambda,\mu)) = f(x) + \langle \lambda, g(x)\rangle + \langle \mu, h(x)\rangle$$

Suppose that $x^\star \in X$ is a local minimum of f on X and suppose that **the constraints are qualified** at $x^\star$, then there exists $\mu^\star \in \mathbb{R}^p$, and $\lambda^\star \in \mathbb{R}^q$ such that :

$$\nabla_x \mathcal{L}(x^\star,(\lambda^\star,\mu^\star)) = 0, \quad \nabla_\mu \mathcal{L}(x^\star,(\lambda^\star,\mu^\star)) = 0, \quad \lambda^\star \succeq 0$$

and for all $i = 1 \dots p$, we must have :

$$\frac{\partial \mathcal{L}}{\partial \lambda_i}(x^\star,(\lambda^\star,\mu^\star)) = 0 \quad \text{or} \quad \lambda_i^\star = 0.$$

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 recall several definition

Definition 7.5.1 Let f, g and h be differentiable functions. We have the following definitions

- The Lagrangian $\mathcal{L}$ is defined by $\mathcal{L}(x,(\lambda,\mu)) = f(x) + \langle \lambda, g(x)\rangle + \langle \mu, h(x)\rangle$
- A point $M = (x,(\lambda,\mu))$ is said to be a KKT point if

$$\nabla_x \mathcal{L}(M) = 0, \quad h(M) = 0, \quad g(M) \preceq 0, \quad \lambda \succeq 0$$

  and for all $i = 1 \dots p$, we have :

$$g_i(M) = 0 \quad \text{or} \quad \lambda_i = 0.$$

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

$$V_M = \left\{ d \in \mathbb{R}^n \text{s.t.} \begin{cases} \langle \nabla h_j(x), d\rangle = 0 & \forall j = 1,\dots,q \\ \langle \nabla g_i(x), d\rangle = 0 & \forall i \text{ s.t. } g_i(x) = 0 \text{ and } \lambda_i > 0 \\ \langle \nabla g_i(x), d\rangle \leq 0 & \forall i \text{ s.t. } g_i(x) = 0 \text{ and } \lambda_i = 0 \end{cases} \right\}$$

Theorem 7.5.3 Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a C^2 function, $g : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^n \to \mathbb{R}^q$ be differentiable functions and let

$$X = \{x \in \mathbb{R}^n \text{ such that } h(x) = 0 \text{ and } g(x) \preceq 0\}.$$

Suppose that the **constraints are qualified** at point $x^\star \in X$

1. If $x^\star \in X$ is a local minimum of f on X then there exists $(\lambda^\star,\mu^\star)$ such that $M^\star = (x^\star,(\lambda^\star,\mu^\star))$ is a KKT point. Moreover,

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

2. If there exists $(\lambda^\star, \mu^\star)$ such that $M^\star = (x^\star, (\lambda^\star, \mu^\star))$ is a KKT point and if

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

then $x^\star$ is a local minimum of f over X.

Exercise 7.7

Write down the theorem equivalent to Theorem 7.5.3 when maximizing f over X and not minimizing. The definition of the Lagrangian must be unchanged, that is

$$\mathcal{L}(x, (\lambda, \mu)) = f(x) + \langle \lambda, g(x)\rangle + \langle \mu, h(x)\rangle$$

Solution of Exercise 7.7

Let $\tilde{\mathcal{L}}$ be the lagrangian for $-f$. The objective is to minimize $-f$, hence Theorem 7.5.3 applies for $\tilde{\mathcal{L}}$. we remark $\tilde{\mathcal{L}}(x, (\lambda, \mu)) = -\mathcal{L}(x, (-\lambda, -\mu))$. We say that $(x, (\lambda, \mu))$ is a KKT point of $\mathcal{L}$ if and only if $(x, (-\lambda, -\mu))$ is a KKT point of $\tilde{\mathcal{L}}$. We have the following definition of a KKT point for a maximization problem : $M = (x, (\lambda, \mu)$ is a KKT point iff

$$\nabla_x \mathcal{L}(M) = 0, \qquad h(M) = 0, \quad g(M) \preceq 0, \quad \lambda \preceq 0$$

and for all $i = 1 \dots p$, we have :

$$g_i(M) = 0 \quad \text{or} \quad \lambda_i = 0.$$

The definition of V_M is transformed into :

$$V_M = \left\{ d \in \mathbb{R}^n \text{s.t.} \begin{cases} \langle \nabla h_j(x), d\rangle = 0 & \forall j = 1, \dots, p \\ \langle \nabla g_i(x), d\rangle = 0 & \forall i \text{ such that } g_i(x) = 0 \text{ and } \lambda_i < 0 \\ \langle \nabla g_i(x), d\rangle \leq 0 & \forall i \text{ s.t. } g_i(x) = 0 \text{ and } \lambda_i = 0 \end{cases} \right\}$$

And the theorem reads :

1. If $x^\star \in X$ is a local maximum of f on X then there exists $(\lambda^\star, \mu^\star)$ such that $M^\star = (x^\star, (\lambda^\star, \mu^\star))$ is a KKT point (with $\lambda^\star \leq 0$). Moreover,

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

2. If there exists $(\lambda^\star, \mu^\star)$ such that $(x^\star, (\lambda^\star, \mu^\star))$ is a KKT point (with $\lambda^\star \leq 0$) and if

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

then $x^\star$ is a local maximum of f over X.

Exercise 7.8

Let $a \in \mathbb{R}^n$, with $a \neq 0$. 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

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

Find every solution of

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

Solution of Exercise 7.8

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

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

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 $(x-a, x-2a)$ which is the family $(\frac{a}{2}+r, -\frac{a}{2}+r)$. Since $(a|r) = 0$, this family is linearly dependent iff $r = 0$ which is in contradiction with $\|r\|^2 = \frac{1}{2}\|a\|^2$.

  The constraints are qualified at each point of X.
- **Step 3 : KKT** The main KKT equation is

$$x - \lambda_1(x-a) + \lambda_2(x-2a) = 0 \tag{7.5}$$

We begin the discussion

- If $\lambda_1 = \lambda_2 = 0$, then (7.5) yields $x = 0$. But $g_2(0) > 0$.

- If $\lambda_1 = 0$ and $\lambda_2 \neq 0$, then (7.5) 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 (7.5) 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 (7.5) 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 empty. 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.In every other case, we cannot conlcude, and we have to compute the value of f at these point and to find that it is the same.

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

♠ 7.6 Proof of KKT

The proof of KKT is quite long and is broken into three pieces. The first one deals with the notion of tangent cone and introduces the qualification of constraints. We say that the constraints are qualified when the tangent cone is easy to compute. In the second part, we use the tangent cone and the Farka's lemma to prove the KKT theorem. In the third part, we discuss about qualification of constraints.

♠ 7.6.1 The tangent cone

Definition 7.6.1 — 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 :

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

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

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

Exercise 7.9

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 of Exercise 7.9

We show that the cone 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 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 $(0,0)$.
We show that $(-1,0)$ does not belongs to $T_{(0,0)}(X)$. 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 -1, which is impossible.

Proposition 7.6.1 $T_x(X)$ 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.

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

$$X = \{x \in \mathbb{R}^n \text{ such that } h(x) = 0, g(x) \preceq 0\},$$

where g and h are vector-valued maps into, respectively $\mathbb{R}^p$ and $\mathbb{R}^q$.

Proposition 7.6.2 Suppose that h and g are differentiable, then for all $d \in T_x(X)$ we have :

- For all $j = 1, \ldots, q$, $\langle \nabla h_j(x), d \rangle = 0$
- For all $i = 1, \ldots, p$, if g_i is active in 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.

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

We then have $h(x_n) = 0$ and $g(x_n) \leq 0$ for all $n \in \mathbb{N}$. By Taylor:

$$h_j(x_n) = h_j(x) + \langle \nabla h_j(x), \varepsilon_n d_n \rangle + o(\varepsilon_n d_n) \text{ and same for } g$$

Using that $h_j(x) = h_j(x_n) = 0$, that $g_i(x_n) \leq 0$ and $g_i(x) = 0$ whenever the constraint is active in x, we obtain :

$$\langle \nabla h_j(x), \varepsilon_n d_n \rangle + o(\varepsilon_n d_n) = 0 \quad \langle \nabla g_i(x), \varepsilon_n d_n \rangle + o(\varepsilon_n d_n) \leq 0$$

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

Definition 7.6.2 — Qualification of constraints. Suppose that h et g are differentiable, we say that the constraints are qualified, when $T_x(X)$ is exactly the set of d that verify :

- For all $j = 1, \ldots, q$, $\langle \nabla h_j(x), d \rangle = 0$
- For all $i = 1, \ldots, p$, if g_i is active in x then $\langle \nabla g_i(x), d \rangle \leq 0$

Exercise 7.10

Show that the constraints are not qualified in the case

$$K = \{(x, y) \text{ such that } (x-1)^2 + y^2 \leq 1 \text{ and } (x+1)^2 + y^2 \leq 1\}.$$

Show that the conclusions of KKT theorem do not apply in this case.

♠ 7.6.2 Proof of first order conditions of KKT

Proposition 7.6.3 Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^\star$ a local minimum of f on X, then

$$\forall d \in T_{x^\star}(X),\ \langle \nabla f(x^\star), d\rangle \geq 0. \tag{7.6}$$

Proof

Let $d \in T_{x^\star}(X)$, introduce $(d_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$ with :

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

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 :

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

Hence: $\varepsilon_n \langle \nabla f(x^\star), d_n\rangle + o(\varepsilon_n d_n) \geq 0$, and dividing by ε_n and letting n goes to $+\infty$ yields the result.

Theorem 7.6.4 — First order KKT with equalities only. Suppose that

$$X = \{x \in \mathbb{R}^n \text{ such that } h(x) = 0$$

and that $x^\star \in X$ is a local minimum of f on X and suppose that "the constraints are qualified", then there exists $\lambda^\star \in \mathbb{R}^q$ such that :

$$\begin{cases} \nabla f(x^\star) + \sum_{j=1}^q \lambda_j^\star \nabla h_j(x^\star) = 0 \\ h(x^\star) = 0 \end{cases}$$

Proof

By qualification of constraints $T_{x^\star}(X) = Vect(\nabla h_j(x^\star))^\perp$. It is a vector space so that upon taking d and $-d$ in (7.6), we have

$$\forall d \in T_{x^\star}(X),\ \langle \nabla f(x^\star), d\rangle = 0.$$

So that

$$\nabla f(x^\star) \subset T_{x^\star}(X)^\perp = Vect(\nabla h_j(x^\star))$$

Hence there exist coefficients $(\kappa_j)_j$ such that $\nabla f(x^\star) = \sum_{j=1}^q \kappa_j \nabla h_j(x^\star)$. Define $\lambda_j^\star = -\kappa_j$ and the proof is complete.

Theorem 7.6.5 — First order KKT with inequalities only. If $x^\star \in X$ is a local minimum of f on X and suppose that "the constraints are qualified", then $\exists \mu^\star \in \mathbb{R}^p$ such that :

$$\begin{cases} \nabla f(x^\star) + \sum_{i=1}^p \mu_i^\star \nabla g_i(x^\star) = 0 \\ g(x^\star) \leq 0, \mu_i^\star g_i(x^\star) = 0, \mu_i^\star \geq 0 \end{cases}$$

Proof

Use Farkas' lemma and

$$T_{x^\star}(X) = \{d \text{ s.t. } \langle \nabla g_i(x^\star), d\rangle \leq 0 \text{ for all } i \text{ such that } g_i \text{ active at } x^\star\}$$

Proposition 7.6.6 — Farkas' lemma. Let $(a_i)_{0\leq i\leq p}$ in $\mathbb{R}^n$. Consider the sets $C = \{d \text{ s.t. } \langle a_i, d\rangle \leq 0 \quad \forall i\}$ and $C^\star = \{w \text{ s.t. } \exists \lambda \geq 0 \text{ s.t. } w = -\sum_i \lambda_i a_i\}$. Then $C^\star$ is a closed convex set and

$$(\langle b, c\rangle \geq 0 \quad \forall c \in C) \text{ iff } b \in C^\star$$

Proof

- Then prove that $C^\star$ is a closed convex set by induction on p, the number of a_i. The case $p = 1$ is trivial. Take $(w_n)_n$ a converging sequence of elements of $C^\star$ towards some w. If the $(a_i)_i$ are linearly independent, then the decomposition $w_n = -\sum_i \lambda_i^n a_i$ is unique and convergence of $(w_n)_n$ is equivalent ot the convergence of $(\lambda^n)_n$ and then $(\lambda^n)_n$ converges to $\lambda \succeq 0$ and $w = -\sum_i \lambda_i^n a_i \in C^\star$. Suppose that the (a_i) are not linearly independent, there exist a linear combination of the form $\sum_i \mu_i a_i = 0$. Take any $w = \sum_i \lambda_i a_i \in C^\star$. Because $\lambda \succeq 0$, there exists a small t such that $\lambda + t\mu \succeq 0$ and for at least one index i_0 we have $\lambda_{i_0} + t\mu_{i_0} = 0$. It follows that

$$w = -\sum_i \lambda_i a_i - t\sum_i \mu_i a_i = -\sum_{i\neq i_0} (\lambda_i + t\mu_i) a_i.$$

Hence we have proven that

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

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

- The implication $\Leftarrow$ is straightforward.
- We prove $\Rightarrow$. Suppose *ad absurdio* that $\exists b$ such that $\langle b, d\rangle \geq 0 \quad \forall d \in C$ but that $b \notin C^\star$. By the theorem of separation of a convex closed set, there exists an $\alpha \in \mathbb{R}$ and a $b^\star \neq 0$ such that

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

Since $0 \in C^\star$, we must have $\alpha < 0$. We prove that $b^\star \in C$, indeed if it were not the case, there would exists an i such that $\langle b^\star, a_i\rangle > 0$ an by choosing $e = -\lambda a_i$ with λ big enough, we have $-\lambda\|a_i\|^2 > \alpha$, which is impossible. Hence $b^\star \in C$ and then we have

$$\langle b, b^\star\rangle < \alpha < 0$$

which is incompatible with our supposition that $\langle b, c\rangle \geq 0 \quad \forall c \in C$.

Farka's lemma relies on the separation of convex set, it is a very important theorem of convexity, originally due to Minkowski. It is extended to vector spaces by Hahn and banach and is known as the Hahn-Banach theorem (geometrical form).

Proposition 7.6.7 Let $C \subset \mathbb{R}^n$ be a convex non-empty set and $b \notin C$, then there exists $b^\star$ and $\alpha \in \mathbb{R}$ such that

$$\langle b^\star, b\rangle < \alpha < \langle b^\star, c\rangle \quad \forall c \in C$$

Proof

Denote π the orthogonal projection of b on C, that is π is the unique minimizer of

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

Note that the problem is strictly convex and that the function is coercive, hence there indeed exists a unique minimizer π. Note also that forall $c \in C$, and forall $\theta \in [0, 1]$, we have $\pi + \theta(c - \pi) \in C$, by convexity of C. Hence $c - \pi$ is an admissible direction at π and we must have

$$\langle \nabla f(\pi), c - \pi\rangle \geq 0.$$

Denote $b^\star = \pi - b = \nabla f(\pi)$. Because $b \notin C$, we have $b^\star \neq 0$ and

$$0 \leq \langle \nabla f(\pi), c - \pi\rangle = \langle b^\star, c - \pi\rangle = \langle b^\star, c - b + b - \pi\rangle = \langle b^\star, x - b - b^\star\rangle$$

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.

♠ 7.6.3 Proving qualification of constraints

The qualificiation of constraints is given in Definition 7.6.2 in an abstract way. In this Section, we give sufficient conditions that ensure constraint qualification.

Proposition 7.6.8 — Qualification of constraints in the equality case. Suppose that X is given by equality constraints only and suppose that the family

$$(\nabla h_j(x^\star))_{j=1,..,q}$$

is linearly independent. Then the constraints are qualified at point $x^\star$, that is :

$$T_x(X) = Vect(\nabla h_j(x^\star))_j^\perp = Ker(Jac[h]_{x^\star}).$$

Proof

We recall that the Jacobian matrix of h is given in an orthonormal basis by :

$$Jac_x[h] = \begin{pmatrix} \nabla h_1(x)^T \\ \dots \\ \nabla h_q(x)^T \end{pmatrix},$$

the matrix $Jac_x[h]h$ is then of size $n \times q$. It is immediate that

$$Vect(\nabla h_j(x))_j^\perp = Ker(Jac_x[h]h).$$

Step 1: Denote $E = Vect(\nabla h_j(x^\star))_j$, $E^\perp = Ker(Jac_{x^\star}[h])$.
Decompose $R^n = E \oplus E^\perp$, and write $x = x_E + x_{E^\perp}$ with $x_E \in E$ and $x_{E^\perp} \in E^\perp$.
Denote $\tilde{h}$:

$$\begin{array}{rcl} \tilde{h} : E \times E^\perp & \to & \mathbb{R}^q \\ (x_E, x_{E^\perp}) & \mapsto & h(x_E + x_{E^\perp}) \end{array}$$

We denote $Jac_{x_E^\star}[\tilde{h}]$ the Jacobian of $\tilde{h}$ with respect to x_E only, it is invertible (because it is **square** and with no kernel).
Use the implicit function theorem with the equation $\tilde{h}(x_E^\star, x_{E^\perp}^\star) = 0$, , there exists neighbourhoods A of $x_E^\star$, and B of $x_{E^\perp}^\star$ and a map $\varphi : B \to A$ such that for all $(x_E, x_{E^\perp}) \in A \times B$:

$$\tilde{h}(x_E, x_{E^\perp}) = 0 \Leftrightarrow x_E = \varphi(x_{E^\perp}).$$

Let $\psi : x_{E^\perp} \mapsto \varphi(x_{E^\perp}) + x_{E^\perp}$. We have $h \circ \psi = 0$. And then

$$Jac_{x^\star}[h].Jac_{x_{E^\perp}^\star}[\psi] = 0.$$

In particular $Im(Jac_{x_{E^\perp}^\star}[\psi]) \subset Ker(Jac_{x^\star}[h])$. Moreover, since

$$Jac_{x_{E^\perp}^\star}[\psi] = \begin{pmatrix} Jac_{v_{E^\perp}^\star}[\phi] \\ I_{n-q} \end{pmatrix}$$

we get $Jac_{x^\star_{E^\perp}}[\psi]$ is at least of rank $(n-q)$.
Since $Ker(Jac_{x^\star}[h])$ is of dimension $(n-q)$, we get

$$Im(Jac_{x^\star_{E^\perp}}[\psi]) = Ker(Jac_{x^\star}[h])$$

Step 2: Let $d \in Ker(Jac_{x^\star}[h])$, there exists $z \in E^\perp$ such that $Jac_{x^\star_{E^\perp}}[\psi].z = d$. Let $(\varepsilon_n)_{n\in\mathbb{N}}$ go to 0. By Taylor

$$\psi(x^\star_{E^\perp} + \varepsilon_n z) = \psi(x^\star_{E^\perp}) + \varepsilon_n Jac_{x^\star_{E^\perp}}[\psi].z + o(\varepsilon_n) = x^\star + \varepsilon_n d + o(\varepsilon_n)\,.$$

If we set $w_n = \psi(x^\star_{E^\perp} + \varepsilon_n z)$, then $h(w_n) = 0$ hence $w_n \in X$ and the sequence $(w_n)_n$ goes to $x^\star$. Moreover

$$w_n = x^\star + \varepsilon_n(d + o(1)) \qquad \text{with } d + o(1) \to d.$$

so that $d \in T_{x^\star}(X)$.

Proposition 7.6.9 — Constraints qualifications for inequalities only. Suppose that X is given by inequality constraints, if there exists a vector u such that forall i such that g_i is active at $x^\star$:

$$\begin{cases} \langle \nabla g_i(x^\star), u\rangle < 0 \\ \text{ or } \\ \langle \nabla g_i(x^\star), u\rangle = 0 \quad \text{and } g_i \text{ is affine} \end{cases}$$

Then the constraints are qualified at point $x^\star$ and the direction u is called a **common re-entry direction**.

For instance, in the important case of Linear programming, all the constraints are affine linear constraints and we can take $u = 0$ as a common re-entry direction. Hence the constraints are always qualified for Linear Programming.

Proof

The vector u is called **common re-entrant direction**. Let $w \in \mathbb{R}^n$ such that for all constraint g_i active at $x^\star$, we have: $\langle \nabla g_i(x^\star), w\rangle \leq 0$. We prove that $w \in T_x(X)$. Let u be a common re-entrant direction, $(\varepsilon_n)_n$ go to zero and let $\eta > 0$. Set :

$$x_n = x^\star + \varepsilon_n(w + \eta u).$$

The sequence $(x_n)_n$ converges towards $x^\star$ and if $x_n \in X$ for n sufficiently large, then $w + \eta u \in T_{x^\star}(X)$ for each η, and since the tangent cone is closed, then w belongs to the tangent cone. Hence we have to prove that $g_i(x_n) \leq 0$ for all i and for n sufficiently large.

1. If g_i is not active at $x^\star$. Then $g_i(x^\star) < 0$ and since x_n converges to $x^\star$, by continuity of g, then $g(x_n) < 0$ for n sufficiently large.
2. If g_i is affine and active at $x^\star$. Then $\langle \nabla g_i(x^\star), u\rangle = 0$ and we have

$$g_i(x_n) = g_i(x^\star) + \varepsilon_n \langle \nabla g_i(x^\star), w + \eta u \rangle = \varepsilon_n \langle \nabla g_i(x^\star), w \rangle \leq 0$$

3. If g_i is active and not affine at $x^\star$. Then $\langle \nabla g_i(x^\star), u \rangle < 0$ and

$$\begin{aligned} g_i(x_n) &= \varepsilon_n \langle \nabla g_i(x^\star), w + \eta u \rangle + o(\varepsilon_n) \\ &= \varepsilon_n \langle \nabla g_i(x^\star), w \rangle + \eta \varepsilon_n \langle \nabla g_i(x^\star), u \rangle + o(\varepsilon_n) \\ &\leq \eta \varepsilon_n \langle \nabla g_i(x^\star), u \rangle + o(\varepsilon_n) \leq 0 \text{ for } n \text{ big enough} \end{aligned}$$

Proposition 7.6.10 — Qualification of constraints, mixed case. Suppose that X is given by equality and inequality constraints, if

- The vectors $(\nabla h_j(x^\star))_{j=1,..,q}$ are linearly independent.
- There exists a vector $u \in Vect((\nabla h_j(x^\star)))_j^\perp$ such that for each g_i active constraint at x :

$$\langle \nabla g_i(x^\star), u \rangle < 0.$$

Then the constraints are qualified.

The proof is an adaptation of the two proofs of the preceding section.

Proposition 7.6.9 is one of the sharpest available, but is sometimes hard to check, we can give weaker forms of Proposition 7.6.9, which is the one that have been studied in Section **??**.

Proposition 7.6.11 — See Proposition 7.4.3. Let $X = \{x, g(x) \preceq 0, h(x) = 0\}$ and let $x^\star \in X$. If the family $\nabla h_j(x^\star)_{j=1,..,q} \cup (\nabla g_i(x^\star))_{i \in I(x^\star)}$ is linearly independent. Then the constraints are qualified at point $x^\star$.

Proof

We exhibit a common re-entry condition. Construct the family $(z_n)_{n=1,...,N}$ such that $z_j = \nabla h_j(x^\star)$ for $j \leq q$ and $z_{q+i} = \nabla g_i(x^\star)$ for $i \in I(x^\star)$. The family $(z_n)_{n=1,...,N}$ is linearly independent. The matrix A whose coefficients are $A_{ij} = \langle \nabla z_i(x^\star), \nabla z_j(x^\star) \rangle$ with $1 \leq i, j \leq N$ is invertible. Then solve $A\alpha = b$ with $b_n = 0$ if $n \leq q$ and -1 if not. Then the vector $u = \sum_n \alpha_n z_n$ verifies verifies $\langle \nabla z_n, u \rangle = b_n$. It is then a common re-entry direction.

Finally, in the case of convex functions, we have the very simple proposition which does not require to compute the gradient of the constraints or to check at different point :

Proposition 7.6.12 — See Proposition 7.4.4. Suppose that X is given by inequality conditions only and that the functions g_i are convex. If there exists a point $x \in X$ such that each constraints which is not affine is inactive at x, then the constraints are qualified everywhere in X.

Proof

Let $x^\star \in X$ and take the direction $u = x - x^\star$. Let $i \in I(x^\star)$ suppose that g_i

is affine, then

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

If g_i is inactive at x then

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

♣ 7.7 Exercise

♣ 7.7.1 Some exercises

Exercise 7.11

Let $f(x, y) = x^4 + y^4 - (x - y)^2$.

1. Compute the gradient and the Hessian of f.
2. Is f convex $\mathbb{R}^n$?
3. Show that f is coercive on $\mathbb{R}^2$.
4. Show that f admits a global minimizer on $\mathbb{R}^2$.
5. Compute the critical points of f.
6. Compute every local minimizer of f. Are they global minimizers ?

Exercise 7.12: Quadratics forms in optimization.

Let $A \in \mathbb{R}^{n \times n}$ be **any** matrix an $b \in \mathbb{R}^n$. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be defined as:

$$f(x) = \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle.$$

1. Compute the gradient and the Hessian of f.
2. If A is symetric, under which conditions is the function f convex ? concave ?
3. *Existence of solutions.* Suppose $A \in \mathbb{R}^{n \times n}$ symetric definite positive
   (a) Let $\lambda_{\max}$ be the largest eigenvalue of A and $\lambda_{\min}$ be the smallest eigenvalue of A. Show that:
   $$\forall x \in \mathbb{R}^n, \quad \lambda_{\min}\|x\|^2 \leq x^\top A x \leq \lambda_{\max}\|x\|^2.$$
   (b) Show that f is coercive.
   (c) Show that f admits a unique global minimum point, denoted $x^\star$, given by $x^\star = A^{-1}b$.
4. Let $g(x_1, x_2) = 2x_1^2 - 2x_1x_2 + x_2^2 - 2x_1$.
   (a) Write g as: $x \mapsto \frac{1}{2}x^\top Ax - b^\top x$ with A symetric
   (b) Give, if they exists, the extrema of g and give their nature.
5. ⋆⋆ **Harder questions :** ⋆⋆
   (a) Write g as : $x \mapsto= \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$. with a non-symetric A.
   (b) What happens if g is changed by
   $$h(x_1, x_2) = -2x_1^2 + 2x_1x_2 - x_2^2 + 2x_1.$$

Exercise 7.13

For each value of β, give the critical points of the function:

$$f(x,y) = x^2 + y^2 + \beta xy + x + 2y$$

and give their nature (min/max, local/global) ?

Exercise 7.14

For each of the following functions, give the local extrema and tell if they are local or global

1. $f(x,y) = (x^2-4)^2 + y^2$.
2. $f(x,y) = (y-x^2)^2 - x^2$.
3. $f(x,y) = \frac{1}{2}x^2 + x\cos(y)$.

Exercise 7.15

We wish to minimize $f : \mathbb{R}^2 \to \mathbb{R}$ on $X \subset \mathbb{R}^2$, where

$$f(x,y) = y - 2x \quad X = \{(x,y) \text{ such that } x^2 + y^2 - 1 \leq 0\}$$

1. Sketch the domain X and the levelsets of f.
2. Show that the problem admits a solution.
3. Show that the problem is convex, what do you conclude?
4. Write down KKT equations and solve them. If you suppose that the constraints are qualified, give the minimizers of f over X.
5. What changes if one wants to maximize f over X ?

Exercise 7.16

Let $A \in \mathcal{M}_{n,n}(\mathbb{R})$ be a symetric definite positive matrix and $b \in \mathbb{R}^n$ with $b \neq 0$. We aim at minimizing f over X, where

$$f(x) = \frac{1}{2}\langle Ax, x\rangle \text{ et } X = \{x \text{ t.q. } \langle b, x\rangle \leq -1\}$$

1. Show that the minimization problem admits a solution, discuss about the maximization problem.
2. Show that the problem is convex, is it strictly convex ?
3. Show that the solution is unique.
4. Write down KKT equations. If you suppose that the constraints are qualified, give the solution of the problem.

Exercise 7.17: Orthogonal projection on a convex set

Let $X \subset \mathbb{R}^n$ be a convex set. The objective is to determine, for every $y \in \mathbb{R}^n$, the orthogonal projection of y on X. It is defined as the unique solution to

$$\min_{x \in X} \frac{1}{2}\|x - y\|^2$$

1. Justify that this problem is strictly convex and admits at most one

solution.
2. Justify that this problem admits at least one solution.
3. If $X = \{\|x\|^2 \leq 1\}$, show that the solution is given by

$$x^\star = \begin{cases} y & \text{si } \|y\| \leq 1 \\ \dfrac{y}{\|y\|} & \text{sinon} \end{cases}$$

4. If X is given by $X = \{g(x) \leq 0\}$ with g convex, and if $x^\star$ is the orthogonal projection of y on X, show that $y - x^\star$ is colinear to $\nabla g(x^\star)$. Justify the use of the term **orthogonal projection**.

♠ 7.7.2 More exercises

Exercise 7.18

Let: $f : (x, y) \in \mathbb{R}^2 \mapsto x^3 + 2xy + y^2$.

1. Is the function f convex ? concave ?
2. Give every local extrema of f and determine if they are local or global.
3. Define the following sets:

$$X_1 = \{(x, y) \in \mathbb{R}^2 \mid y - x < 0\} \quad \text{et} \quad \widetilde{X}_1 = \{(x, y) \in \mathbb{R}^2 \mid y - x > 0\}.$$

Give every local extrema of f on X_1, and then on $\widetilde{X}_1$.

Exercise 7.19

Denote $X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 2\}$ and consider the following problems

$$(P_1) \qquad \min_{(x,y)\in X} f_1(x, y) = x + y.$$
$$(P_2) \qquad \min_{(x,y)\in X} f_2(x, y) = xy.$$

1. Show that the problems (P_1) et (P_2) admit solutions.
2. Show that the constraint $x^2 + y^2 = 2$ is qualified in any point of X.
3. Determine every local extrema of f_1 sur X. Are they global ?
4. Same question for f_2.

Exercise 7.20

Compute the KKT points of

$$(P) \qquad \min_{(x,y)\in\mathbb{R}^2} f(x, y) = -\frac{1}{2}(x - 4)^2 + 2y^2 \quad \text{sous} \quad 1 - x^2 - y^2 \leq 0.$$

Compute the minimizers of (P) and determine if they are local or global.

Exercise 7.21

Let

$$X = \{(x, y, z) \in \mathbb{R}^3 \mid x + y + z = 1,\ x^2 + y^2 + z^2 \leq 1\}.$$

Give the extrema of the function $f : (x, y, z) \in \mathbb{R}^3 \mapsto x$ on X.

Exercise 7.22

Let A be a real $m \times n$ matrix and $b \in \mathbb{R}^n$. Let:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x\|^2 + \langle b, x \rangle \qquad \text{sous: } Ax = 0.$$

1. Show that this problem is equivalent to a projection problem on a certain convex set. Make explicit this convex set
2. Show that if A is of full rank, then the optimal solution is given by:

$$x^* = (A^\top (AA^\top)^{-1} A - I_n)b.$$

Exercise 7.23

Let A be a square matrix of size n. Solve the following problem:

$$\min_{X \in \mathbb{M}_n(\mathbb{R})} f(X) = \frac{1}{2}\|X - A\|_F^2 \qquad \text{sous: } X^\top = X,$$

in the case where the matrix A is (a) symmetric and then (b) non-symmetric.

Exercise 7.24

Let $f(x, y) = y - 2x$ and:

$$X = \{(x, y) \in \mathbb{R}^2 \mid y \geq x^2 \text{ et } x^2 + y^2 \leq 1\}.$$

1. Give the extremal points of f on X. Determine if they are local or global.
2. What happens if we replace $y \geq x^2$ by $y \leq x^2$?

CHAPTER 8

Duality

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

$$(P) \quad \begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.c.:} & g_j(x) \leq 0,\ j = 1,\dots,q\ , \\ & h_i(x) = 0,\ i = 1,\dots,p \end{array}$$

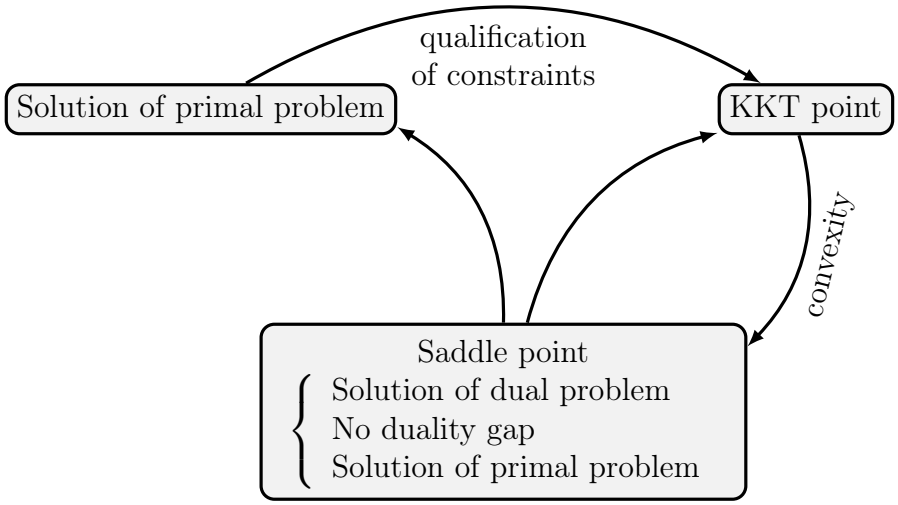
where the functions f, g_j and h_i are C^1 and real-valued. According to KKT theorem, any solution $x^\star$ to such a problem can be associated to Lagrange multiplier $(\lambda^\star, \mu^\star)$. There exists two important families of algorithms that aim at solving the problem: the primal methods and the dual methods. The primal methods aim at finding a point $x^\star$, the Lagrange multiplier $(\lambda^\star, \mu^\star)$ 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 $(\lambda^\star, \mu^\star)$ 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.

On the theoritical side, most of the results that are going to be presented in this chapter can be sum up in the following diagramm, which meaning will be made clear at the end of the chapter :



♠ 8.1 Min-Max 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** (P) and can be written as :

$$(P) \quad \inf_{x \in X} \sup_{y \in Y} \varphi(x, y) \qquad \text{Primal problem}$$

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

$$(D) \quad \sup_{y \in Y} \inf_{x \in X} \varphi(x, y) \qquad \text{Dual problem}$$

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 8.1.1 — Weak duality. Let X, Y be 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

Denote $\Phi_1(y) = \inf_{\tilde{x}\in X} \varphi(\tilde{x}, y)$ and $\Phi_2(x) = \sup_{\tilde{y}\in Y} \varphi(x, \tilde{y})$. 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 $\Phi_1(y) \leq \Phi_2(x)$, 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 8.1.1 — 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 sadlle point allows us to caracterize when the dual problem and the primal problem are the same.

Theorem 8.1.2 — 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 (P).
ii. $\bar{y}$ is a solution of the dual problem (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 $\bar{x}$. 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.

It is possible to caracterize the set of saddle points

Corollary 8.1.3 Let $\varphi : X \times Y \to \mathbb{R}$. Suppose that φ has at least one saddle point, then:

1. The set of saddle points of φ is a cartesian product $\bar{X} \times \bar{Y}$ with $\bar{X} \subset X$ and $\bar{Y} \subset Y$.
2. φ has a constant value denoted $\bar{\varphi}$ on $\bar{X} \times \bar{Y}$.
3. $$\bar{X} = \bigcap_{y \in Y} \{x \in X \mid \varphi(x, y) \leqslant \bar{\varphi}\}$$
   $$\bar{Y} = \bigcap_{x \in X} \{y \in Y \mid \varphi(x, y) \geqslant \bar{\varphi}\}$$

Proof

Todo.

♠ 8.2 Standard form and duality

In this section, we discuss the dual problem of a problem in standard form

Definition 8.2.1 Consider the problem $\inf_{x \in X} f(x)$ where

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i = 1, \dots, p,\ h_j(x) = 0,\ j = 1, \dots, q\}.$$

Introduce the Lagrangian $\mathcal{L}$ defined by

$$\mathcal{L}(x, (\lambda, \mu)) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x) + \sum_{j=1}^{q} \mu_j h_j(x) = f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle.$$

Define the **primal function** $\bar{f}(x) = \min_{\lambda \succeq 0, \mu} \mathcal{L}(x, (\lambda, \mu))$ 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

$$\min_{x \in \mathbb{R}^n} \bar{f}(x).$$

Proof
TODO

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

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

- The **dual admissible domain**:

$$X^\star = \left\{(\lambda,\mu)\in\mathbb{R}^p\times(\mathbb{R}_+)^q : \inf_{x\in\mathbb{R}^n}\mathcal{L}(x,(\lambda,\mu)) > -\infty\right\},$$

  is the set on which the dual function is finite

Proposition 8.2.1 The dual function $f^\star$ is always concave and the admissible domain $X^\star$ is always convex.

Exercise 8.1
Prove it

We already know that if $(\bar{x},(\bar{\lambda},\bar{\mu}))$ is a saddle-point of the Lagrangian, then $\bar{x}$ is a solution of the minimization problem. An important feature of saddle point is that they verify KKT equations. Indeed if $(\bar{x},(\bar{\lambda},\bar{\mu}))$ is a saddle-point, then

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

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

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

Which gives immediatly the KKT conditions:

$$\nabla f(\bar{x}) + \sum_{i=1}^{p}\lambda_i\nabla g_i(\bar{x}) + \sum_{j=1}^{q}\mu_j\nabla h_j(\bar{x}) = 0.$$

Theorem 8.2.2 — Saddle points are KKT. Consider the problem

$$\inf_{x\in\mathbb{R}^n} f(x) \quad \text{s.c.} \quad \begin{aligned} g_i(x) &\leq 0, \quad i=1,\dots,p,\\ h_j(x) &= 0, \quad j=1,\dots,q,\end{aligned}$$

If a point $(x,(\lambda,\mu))$ is a saddle point of the Lagrangian

$$\mathcal{L}(x,(\lambda,\mu)) = f(x) + \langle\lambda,g(x)\rangle + \langle\mu,h(x)\rangle$$

Then it verifies KKT conditions $\nabla_x\mathcal{L}(x,(\lambda,\mu)) = 0$

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

Theorem 8.2.3 — Convex Lagrangian duality. Consider the problem

$$\inf_{x\in\mathbb{R}^n} f(x) \quad \text{s.c.} \quad \begin{aligned} g_i(x) &\leq 0, \quad i = 1,\dots,p, \\ h_j(x) &= 0, \quad j = 1,\dots,q, \end{aligned}$$

where the functions g_j, $j = 1,\dots,q$, and f are convex and the functions h_i, $i = 1,\dots,p$, are affines. A point $(x,(\lambda,\mu))$ is a saddle point of the Lagrangian

$$\mathcal{L}(x,(\lambda,\mu)) = f(x) + \langle\lambda, g(x)\rangle + \langle\mu, h(x)\rangle$$

if and only if it verifies KKT conditions $\nabla_x\mathcal{L}(x,(\lambda,\mu)) = 0$.

As a corollary, if the constraints are qualified and if x is a minimum of the primal problem, then there exists (λ,μ) such that $(x,(\lambda,\mu))$ is a saddle point.

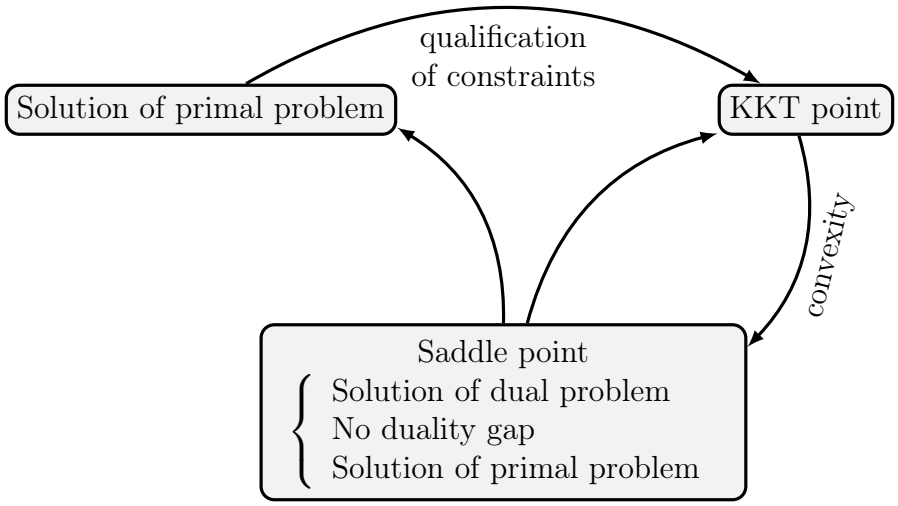
Proof

We already know that saddle points verify the KKt conditions. Let $(\bar{x},(\bar{\lambda},\bar{\mu}))$ be a KKT point of (P), then

$$\nabla f(\bar{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\bar{x}) = 0 \quad \text{soit:} \quad \nabla_x\mathcal{L}(\bar{x}(\bar{\lambda},\bar{\mu})) = 0.$$

Hence $\bar{x}$ is a critical point of $\psi : x \in \mathbb{R}^n \mapsto \mathcal{L}(x,(\bar{\lambda},\bar{\mu}))$. But ψ is a positive linear combination of convex function hence is convex, and $\bar{x}$ is a global minimum of ψ. Moreoever, for every $\lambda \succeq 0$ and μ we have $\mathcal{L}(\bar{x},(\lambda,\mu)) = f(\bar{x}) + \langle\lambda, g(\bar{x})\rangle \geq f(\bar{x}) = \mathcal{L}(\bar{x},(\lambda,\mu))$. Hence $(\bar{x},(\bar{\lambda},\bar{\mu}))$ is a saddle point of the Lagrangian.

Finally, in order to conclude this "tour of duality", we recall the diagramm we showed at the beginning of the section that sums up the relations between the different notions that we saw.



♠ 8.3 Duality of Linear Programming

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 there is no admissible points.

Theorem 8.3.1 There exists a solution if and only if the optimal value is finite

$$-\infty < \inf_{Ax \succeq b, A'x=b'} \langle c, x\rangle < +\infty$$

The optimal value is $+\infty$ if and only if the corresponding

Proof

First put the problem in **equality form** $\inf_{A'x=b'} c$. 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 $z_k = \mathcal{A}x_k$. 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 Farka's lemma, we kno tha $\mathcal{C}$ is closed, there exists some $\lambda \succeq 0$ such that $\mathcal{A}\lambda = \begin{pmatrix} \alpha \\ b \end{pmatrix}$. The minimum is then attained at λ.

The Lagrangian is given by $\mathcal{L}(x, y) = \langle c, x\rangle + \langle y, b - Ax\rangle$. Note that the constraint $x \succeq 0$ is not in the Lagrangian. The original problem is then

- **Inequality constraints** $Ax \preceq b$:

$$\inf_{Ax \preceq b, x \succeq 0} \langle c, x\rangle = \inf_{x \succeq 0} \sup_{y \preceq 0} \mathcal{L}(x, y)$$

- **Equality constraints** $Ax = b$:

$$\inf_{Ax = b, x \succeq 0} \langle c, x\rangle = \inf_{x \succeq 0} \sup_{y} \mathcal{L}(x, y)$$

In any case the dual function is given by

$$f^\star(y) = \inf_{x \succeq 0} \mathcal{L}(x, y) = \inf_{x \succeq 0} \langle y, b\rangle + \langle -A^T y + c, x\rangle = \begin{cases} \langle y, b\rangle & \text{if } c - A^T y \succeq 0 \\ -\infty & \text{otherwise} \end{cases}$$

We deduce the following proposition

Proposition 8.3.2 The dual problem of $\min_{Ax \preceq b, x \succeq 0} \langle c, x \rangle$ is $\max_{A^T y \preceq c, y \preceq 0} \langle b, y \rangle$. The dual problem of $\min_{Ax=b, x \succeq 0} \langle c, x \rangle$ is $\max_{A^T y \preceq c} \langle b, y \rangle$.

The rule is the following. Denote a_i the rows of A and $a_j^\star$ the columns of A. 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 $\mathbb{R}^m$.

Minimize $\langle x, c \rangle$	$\iff$	Maximize $\langle y, b \rangle$
$\langle a_i, x \rangle \geq b_i$		$y_i \geq 0$
$\langle a_i, x \rangle = b_i$	$\iff$	$y_i \in \mathbb{R}$
$\langle a_i, x \rangle \leq b_i$		$y_i \leq 0$
$x_j \geq 0$		$\langle a_j^\star, y \rangle \leq c_j$
$x_j \in \mathbb{R}$	$\iff$	$\langle a_j^\star, y \rangle = c_j$
$x_j \leq 0$		$\langle a_j^\star, y \rangle \geq c_j$

Exercise 8.2

Construct the dual of the following problem:

$$\begin{array}{rrcrcr} \max z = & 4x_1 & + & 5x_2 & & \\ & 2x_1 & + & x_2 & \leq & 800 \\ & x_1 & + & 2x_2 & \leq & 700 \\ & & & x_2 & \leq & 300 \\ & x_1 & , & x_2 & \geq & 0. \end{array}$$

Theorem 8.3.3 Consider a Linear Programming in equality or inequality form.

- If one of the problems (primal or dual) has finite optimal value, or if they both have non-empty admissible points, 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, we have $\langle c, x \rangle \geq \langle y, b \rangle$. We have equality if and only if (x, y) is a saddle point of the Lagrangian.

Proof

First note that a linear programming problem is convex, so that saddle points are exactly KKT points. Second, since all the constraints are affine, the constraints are qualified so that a solution of the primal (or the dual) is a KKT point.

First suppose that both problems have non-empty admissible sets, let x and y be two admissible points, then $c \succeq A^T y$ and $x \succeq 0$ so that

$$\langle c, x \rangle \geq \langle A^T y, x \rangle = \langle y, Ax \rangle \overset{(1)}{\geq} \langle y, b \rangle. \tag{8.1}$$

The inequality (1) is proven either by using $Ax = b$ in the equality case or $-Ax \succeq -b$ and $y \succeq 0$ in the equality case. Inequality (8.1) proves that the primal and dual problems have finite optimal values, hence admit solutions. This proves the third item. We now show that equality in (8.1) proves that we

have a saddle point. To that end suppose that $(\bar{x}, \bar{y})$ is admissible and that

$$\langle c, \bar{x} \rangle = \langle \bar{y}, b \rangle.$$

For all y admissible, (8.1) 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 (8.1) 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 (8.1). The last thing to show is that if there is a finite optimal value to a problem, there is a finite optimal value to the other one. We suppose that we have finite optimal value of the primal problem, hence we have a solution, say $\bar{x}$. Because the constraints are qualified, it is a KKT point and hence there exists $\bar{y}$, Lagrange multiplier. But the problem is convex, hence $(\bar{x}, \bar{y})$ is a saddle point and hence there exists $\bar{y}$ a solution of the dual.

♠ 8.4 Exercises

Exercise 8.3

Consider the following problem

$$(P_1) \quad \left| \begin{array}{ll} \min\limits_{(x,y)\in\mathbb{R}^2} & f(x,y) = 2x + y \\ \text{s.c.} & x^2 + y^2 \leq 1 \end{array} \right.$$

1. Compute $(P_1^\star)$, the dual problem of (P_1).
2. Solve $(P_1^\star)$.

Exercise 8.4

Solve the following problem by using a dual approach

$$\min_{x\in\mathbb{R}^n} \langle a, x \rangle \quad \text{with } \|x\|_2^2 \leq 1.$$

Exercise 8.5

Let

$$(P) \quad \left| \begin{array}{ll} \min\limits_{(x,y)\in\mathbb{R}^2} & -x^2 + y^2 \\ \text{s.c.} & x + y \geq 0 \\ & y \geq x \end{array} \right.$$

1. Draw and redefine the admissible domain of (P). Show very simply that problem (P) admits an infinity of solutions. Give the set of solutions.

2. Put the problem in standard form and show that the associated Lagrangian does not admit saddle points.

Exercise 8.6

Let

$$(P) \quad \left| \begin{array}{ll} \min\limits_{(x,y,z)\in\mathbb{R}^3} & x-y \\ \text{sous:} & x+y+z \leq 1, \quad x^2+y^2+z^2 \leq 1 \end{array} \right.$$

1. Show that problem (P) has a solution.
2. Solve problem (P) by a dual approach.

Exercise 8.7: additional exercise

Consider the optimization problem

$$(P_3) \quad \left| \begin{array}{ll} \min\limits_{(x,y)\in\mathbb{R}^2} & x^2+y^2-x \\ \text{s.t.} & x+y \geq 1 \end{array} \right.$$

1. Put the problem in standard form and give $(P_3^\star)$, its dual.
2. Solve $(P_3^\star)$.
3. Give the solutions of (P_3).

Exercise 8.8: Quadratic optimization

Let A be a square real symetric and invertible matrix of size n. Let $b \in \mathbb{R}^n$ and C be a real matrix of size $p \times n$ and of full rank ($p < n$). Let $d \in \mathbb{R}^p$. We consider the following problem:

$$(P_3) \quad \left| \begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) = \dfrac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle \\ \text{s.c.} & Cx - d = 0 \end{array} \right.$$

1. Under which conditions is f_3^*, the dual function associated to the problem (P_3), well defined ? Compute it.
2. Solve the dual problem (P_3^*). Do we have existence and uniqueness of a solution of (P_3) ? Give the analytical expression of this solution and of the associated multiplicator
3. Replace the equality constraint by an inequality $Cx - d \leq 0$. Write the dual problem associated to the new constraint.

IV Algorithmics

CHAPTER 9

Descent methods for unconstrained smooth optimization

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

♣ 9.1 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 of descent**.
- 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**.

♣ 9.1.1 Direction of descent

Definition 9.1.1 — 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.

♣ 9.2 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 **??**. 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

$$\|\nabla f(x_k)\| < \varepsilon, \tag{9.1}$$

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

In practice, the algorithm may fail to satisfy the test (9.1), 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

$$\text{maximum number of iteration } + (9.1) + \text{ if needed, stagnation criteria}$$

Remark 9.2.1 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.

♣ 9.3 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 9.3.1 — 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**.

Remark 9.3.1 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 9.3.2 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:

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

♣ 9.4 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.

Input: .
Output: .

$s_k = s_{k-1}$

OPTIMAL STEP LINESEARCH.

Input: .
Output: .

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.

Input: .
Output: .

$s_k = s_{k-1}$
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.

Input: .
Output: .

$s_k = 1.3 * s_{k-1}$
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.

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

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.

♠ 9.5 Wolfe line search

♠ 9.5.1 Presentation of Wolfe linesearch

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

Exercise 9.1: 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 $x_0 \neq 0$.

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 of Exercise 9.1

We have $x_{k+1} = x_k + s_k d_k$. In the first case, we have $x_{k+1} = x_k - 2 - 32^{-k}$

$$x_k = (-1)^k(1 + \frac{1}{2^k}).$$

Pour tout $k \in \mathbb{N}$: $f(x_{k+1}) < f(x_k)$: on a donc bien un algorithme de descente mais la suite $(x_k)_{k\in\mathbb{N}}$ ne converge pas : elle possède deux points d'accumulation en $x = 1$ et $x = -1$ et aucun de ces deux points n'est un extremum de f.

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 9.5.1 — 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 :**
$$f(x_k + s_k d_k) < f(x_k) + \varepsilon_1 s_k \langle \nabla f(x_k), d_k \rangle \tag{9.2}$$
- **Avoid small steps :**
$$\langle \nabla f(x_k + s_k d_k), d \rangle > \varepsilon_2 \langle \nabla f(x_k), d_k \rangle \tag{9.3}$$

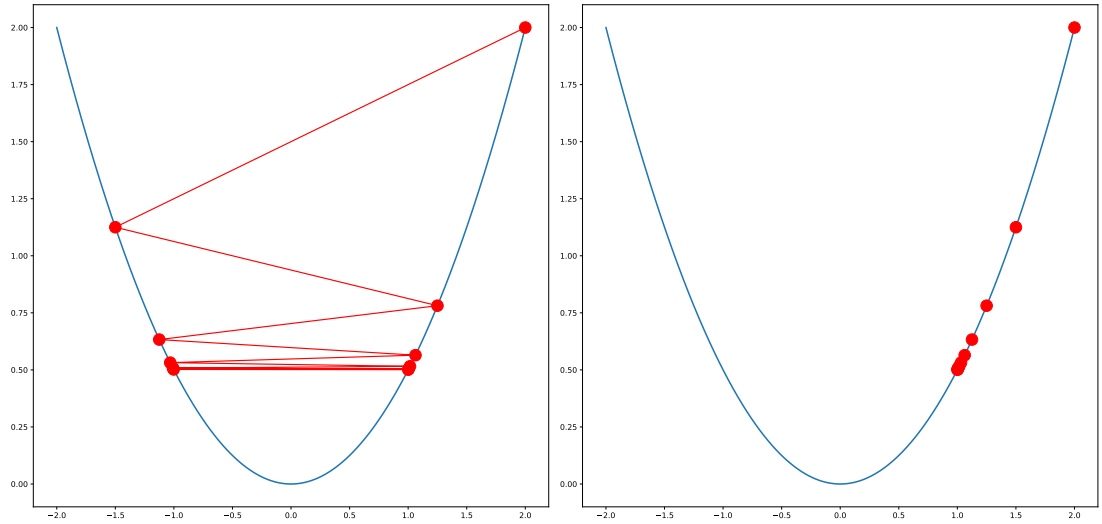


Figure 9.1: Examples of big and small steps

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 $\varepsilon_2 = 0.99$.

Remark 9.5.1 Notation with the merit function In the proof, we introduce the following notation :

$$\phi : s \mapsto f(x_k + sd_k) - f(x_k),$$

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

$$\phi'(s) > \varepsilon_2 \phi'(0) \text{ and } \phi(s) < \varepsilon_1 s \phi'(0)$$

♠ 9.5.2 Computation of a Wolfe step

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

Proposition 9.5.2 — 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 $\varepsilon_1 < 1$, $\mathcal{A} \neq \emptyset$. Denote $s = \sup \mathcal{A}$. Since f is bounded from below, $s \neq +\infty$. By continuity of f, 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

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

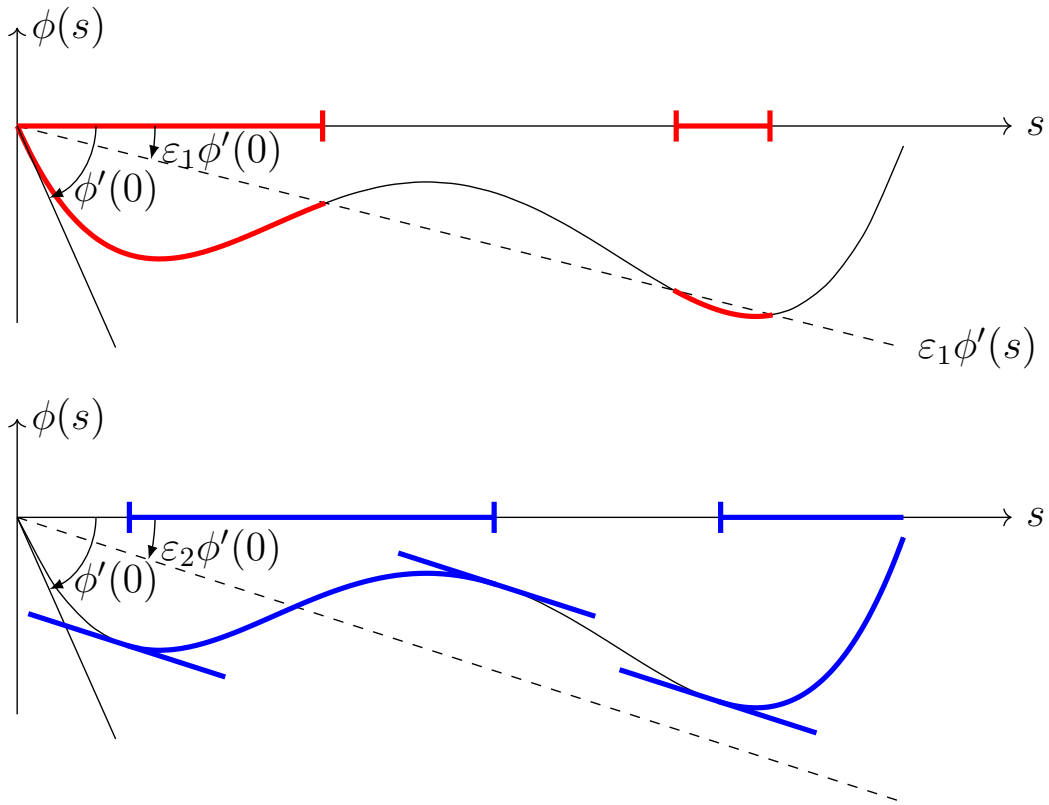


Figure 9.2: The Wolfe conditions. Top, the steps s such that $\phi(s) < \varepsilon_1 s\phi'(0)$. Bottom, the steps such that $\phi'(s) > \varepsilon_2\phi'(0)$.

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 $(u_n - s_k)$, and let n goes 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 (9.2):

$$s_+ = s_k \quad \text{and} \quad s_{k+1} = \frac{s_- + s_+}{2}.$$

(b) **Too small** If s_k meets (9.2) but not (9.3) :

$$s_- = s_k \quad \text{and} \quad s_{k+1} = \begin{cases} \dfrac{s_- + s_+}{2} & \text{if } s_+ < +\infty \\ 2s_k & \text{else.} \end{cases}$$

(c) $k := k + 1$;

3. Return $s^\star = s_k$.

We admit without proof that under the exact same hypothesis of Proposition 9.5.2, 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.

♠ 9.5.3 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 $\nabla f(\bar{x}) = 0$).

Remark 9.5.3 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 9.5.4 — 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 $\cos(\theta_k) = \frac{\langle -\nabla f(x_k), d_k\rangle}{\|\nabla f(v_k)\|\|d_k\|}$, 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

$$s_k \geq \frac{\varepsilon_2 - 1}{L} \frac{\langle \nabla J(v_k), d_k \rangle}{\|d_k\|^2} > 0.$$

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 :

$$f(x_0) - \inf_{x \in \mathbb{R}^n} f(x) \geq C \sum_{k \geq 0} \cos(\theta_k)^2 \|\nabla f(x_k)\|^2.$$

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 9.5.5 As a corollary of Theorem 9.6.3, 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

$$K = \frac{L}{c^2 \varepsilon_1 (1 - \varepsilon_2)} \left(f(x_0) - \inf_{x \in \mathbb{R}^n} f(x) \right)$$

♣ 9.6 Gradient algorithms

♣ 9.6.1 Description algorithm

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

Proposition 9.6.1 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

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

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 9.6.1 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.

♣ 9.6.2 Convergence of Gradient algorithm

Theorem 9.6.2 — 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 $x_{k+1} = x_k - s_k \nabla f(x_k)$.

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle| + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - s_k\|\nabla f(x_k)\|^2 + \frac{L}{2}s_k^2\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) + s_k\left(\frac{L}{2}s_k - 1\right)\|\nabla f(x_k)\|^2. \end{aligned} \tag{9.4}$$

- 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 (9.4) 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 (9.4) 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 9.6.3 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 9.6.4 As a application of Theorem 9.6.3, suppose that the hypothesis of Theorem 9.6.2 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.

♣ 9.6.3 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.

step	0.325	0.25	0.125	0.05	0.01
iteration number	DV	49	101	263	1340

Table 9.1: Number of iterations of the fixed step gradient algorithm in order to approach a critical point of f within 10^{-5} accuracy. Initial point $x_0 = (7, 1.5)$.

- **Optimal step strategy :** At each iteration the optimal step strategy amounts to solve

$$\min_{s>0} f(M_k + sd_k) = frac12x_k^2(1-s)^2 + \frac{7}{2}y_k^2(1-7s)^2$$

The above function is a second order polynom in s, the solution of the above problem is given by:

$$s_k = \frac{x_k^2 + 7^2y_k^2}{x_k^2 + 7^3y_k^2}.$$

In order to reach a critical point with 10^{-5} accuracy, starting with $x_0 = (7, 1.5)$, the algorithm requires 43 iterations.

On Figure 9.6.3, 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.

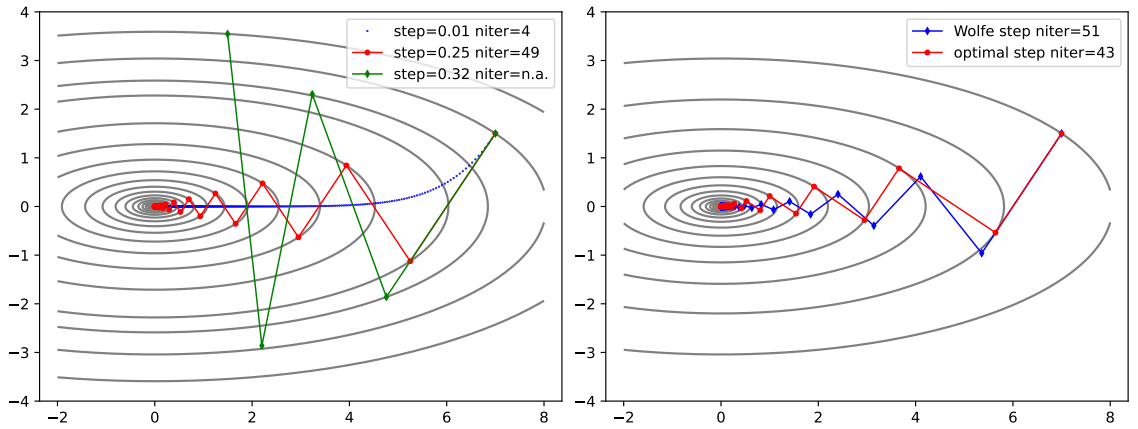


Figure 9.3: Gradient algorithm for a quadratic function with initial point $x_0 = (7, 1.5)$. On the Left, fixed step algorithm and on the right : Steepest descent algorithm (red) and Wolfe linesearch algorithm (blue).

♣ 9.7 Newton algorithms

♣ 9.7.1 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 9.7.1 — 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 9.7.1 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.

♣ 9.7.2 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 9.7.2 — 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.

♣ 9.7.3 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.

♣ 9.7.4 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 !!!

♣ 9.7.5 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 :

$$\begin{array}{ll}((-1)^{k+1}, k\pi),\ k\in\mathbb{Z} & \text{local minima of } g\\ (0, \frac{\pi}{2}+k\pi),\ k\in\mathbb{Z} & \text{saddle-points of } g\end{array}$$

In Figure 9.6, we show the results for the optimization of the Rosenbrock function and in Figure 9.7 for the Oscill function.

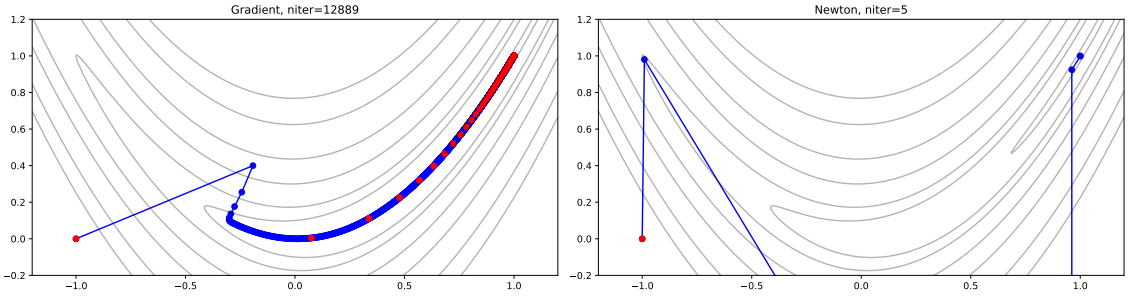


Figure 9.4: Rosenbrock:

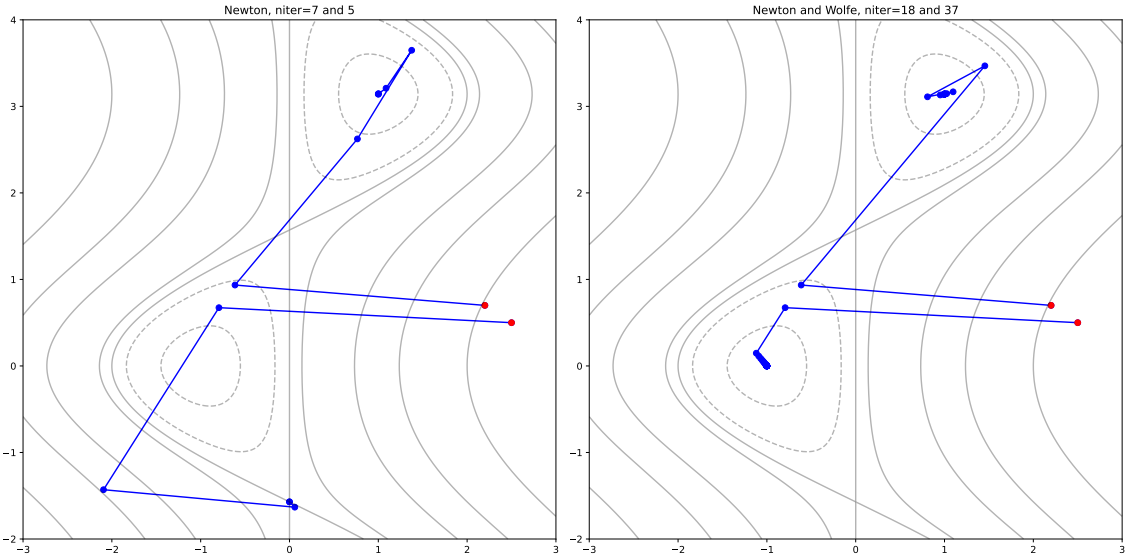


Figure 9.5: Oscill:

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.

♠ 9.7.6 Convergence of Newton

The Newton's method enjoys quadratic convergence, the result is made precise in the following theorem.

Theorem 9.7.2 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

Exercise 9.2

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 9.7.2 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 9.7.3 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 9.7.4 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 9.6, we show the results for the optimization of the Rosenbrock function and in Figure 9.7 for the Oscill function.

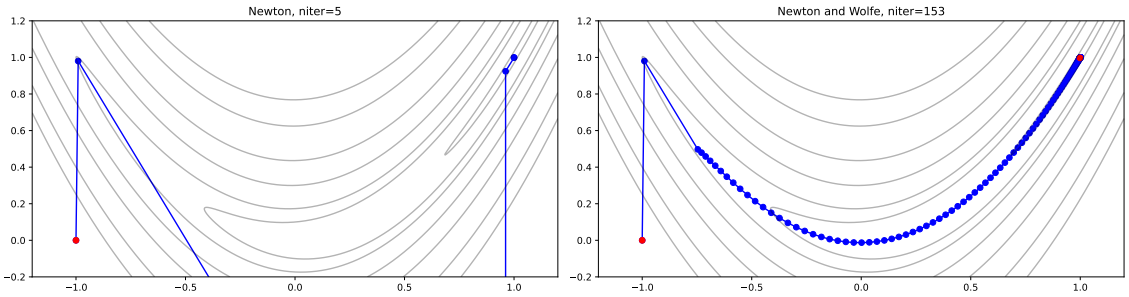


Figure 9.6: Rosenbrock:

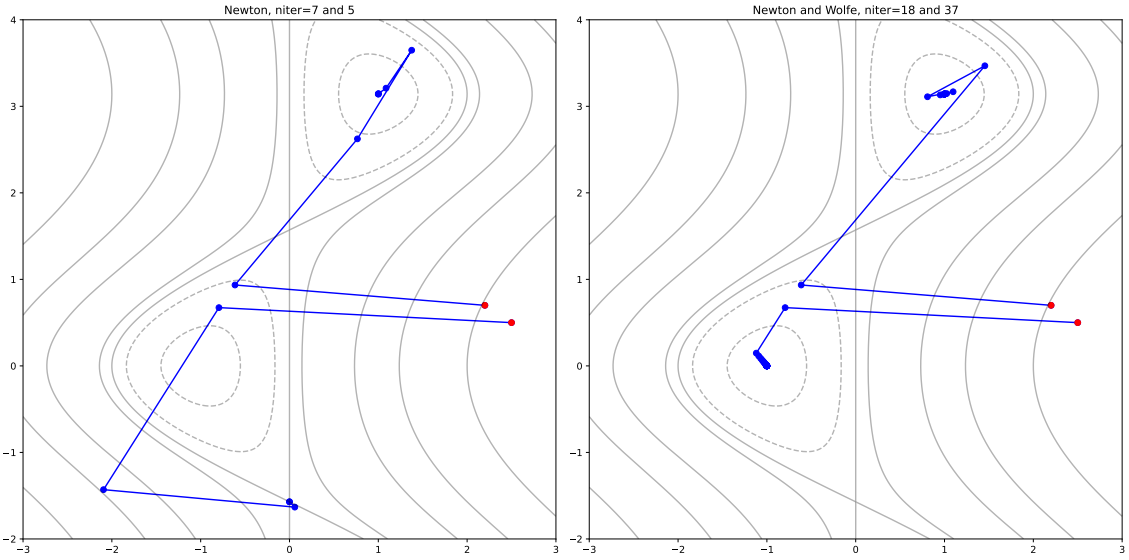


Figure 9.7: Oscill:

♣ 9.8 Quasi-Newton algorithm

♣ 9.8.1 Defintion

The class of **Quasi-Newton** algorithms is a class of algorithms such that the direction of descent d_k is computed by

$$d_k = H_k^{-1}(-\nabla f(x_k)),$$

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.

♣ 9.8.2 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 :

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2} \sum_{i=1}^{m} F_i(x)^2, \tag{9.5}$$

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 9.8.1 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^TA is symetric and $A^TA \succeq 0$, because for every vector h, we have

$$\langle A^TAh, h\rangle = \langle Ah, Ah\rangle = \|Ah\|^2$$

♠ 9.8.3 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 9.8.2 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 $x^\star = \left(A^TA\right)^{-1}A^Tb$

♠ 9.8.4 The BFGS Algorithm

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

Definition 9.8.1 The BFGS class of algorithms are Quasi-Newton algorithms with direction of descent given by $d_k = -H_k^{-1}\nabla f(x_k)$, where H_k is constructed so that H_k is symmetric, positive definite and at each iteration

$$H_{k+1}\sigma_k = y_k \text{ if } \sigma_k = x_{k+1} - x_k \text{ and } y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

A necessary condition for a BFGS algorithm to be built is $(\sigma_k, y_k) > 0$.

Proof

If there exists a positive definite matrix such that $H_{k+1}\sigma_k = y_k$, then we must have $(H_{k+1}\sigma_k, \sigma_k) > 0$. Hence a necessary condition for the existence of a BFGS algorithm is $(\sigma_k, y_k) > 0$.

Definition 9.8.2 — DFP-Method (Davidson, Fletcher, Powell. 1959-63). The DFP update is to set H_k as the solution of

$$\min_{H=H^T, H\sigma_k=y_k} \|H - H_k\|_W,$$

for a well chosen norm, this leads to the formula, if $B_k = H_k^{-1}$.

$$H_{k+1} = \left(I - \frac{y_k\sigma_k^T}{\langle y_k, \sigma_k\rangle}\right) H_k \left(I - \frac{\sigma_k y_k^T}{\langle y_k, \sigma_k\rangle}\right) + \frac{y_k y_k^T}{\langle y_k, \sigma_k\rangle}.$$

$$B_{k+1} = B_k + \frac{\sigma_k\sigma_k^T}{\langle y_k, \sigma_k\rangle} - \frac{B_k y_k y_k^T B_k}{\langle y_k, B_k y_k\rangle}.$$

Definition 9.8.3 BFGS-Method (Broyden, Fletcher, Goldfarb, Shannon. 1969-70) The BFGS update amounts to work directly with $B_k = H_k^{-1}$ and to set B_{k+1} as the solution of

$$\min_{B=B^T, By_k=\sigma_k} \|B - B_k\|_W,$$

for a well chosen norm, this leads to the formulas

$$B_{k+1} = \left(I - \frac{\sigma_k y_k^T}{\langle y_k, \sigma_k\rangle}\right) B_k \left(I - \frac{y_k\sigma_k^T}{\langle y_k, \sigma_k\rangle}\right) + \frac{\sigma_k\sigma_k^T}{\langle y_k, \sigma_k\rangle}. \text{ with } H_k = B_k^{-1}$$

$$H_{k+1} = H_k + \frac{y_k y_k^T}{\langle y_k, \sigma_k\rangle} - \frac{H_k\sigma_k\sigma_k^T H_k}{\langle \sigma_k, H_k\sigma_k\rangle}.$$

These two methods are duals since we always have $H_k = B_k^{-1}$. We can check that those matrices verify :

- The matrices H_k and B_k are symetric
- We always have $H_{k+1}\sigma_k = y_k$ and $B_{k+1}y_k = \sigma_k$.
- The matices H_k and B_k are always positive definite.

Proposition 9.8.3 If B_k is positive definite and if $\langle \sigma_k, y_k\rangle > 0$, then B_{k+1} is positive definite.

Proof

Denote C as $C = I - \frac{y_k \sigma_k^T}{\langle y_k, \sigma_k \rangle}$ then for any $x \neq 0$ and $u = Cx$,

$$\langle B_{k+1}x, x\rangle = \langle C^T B_k C x, x\rangle + \frac{\langle \sigma_k \sigma_k^T x, x\rangle}{\langle y_k, \sigma_k\rangle} = \langle B_k u, u\rangle + \frac{\langle \sigma_k, x\rangle^2}{\langle y_k, \sigma_k\rangle},$$

Since $\langle \sigma_k, y_k\rangle > 0$, we have $\langle B_{k+1}x, x\rangle \geq 0$ for all $x \in \mathbb{R}^n$. It remains to show that $\langle B_{k+1}x, x\rangle \neq 0$. Since $B_k > 0$, we have

$$\begin{aligned} \langle B_{k+1}x, x\rangle = 0 &\Rightarrow \langle B_k u, u\rangle \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \\ &\Rightarrow u = 0 \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \end{aligned}$$

but $u = Cx = x - \frac{y_k \langle x, \sigma_k\rangle}{\langle y_k, \sigma_k\rangle}$. Hence

$$u = 0 \quad \text{and} \quad \langle x, \sigma_k\rangle = 0 \Rightarrow x = 0$$

Proposition 9.8.4 If we use a Wolfe algorithm, then $\langle \sigma_k, y_k\rangle > 0$, and B_k is positive definite.

Proof

$$\begin{aligned} \langle y_k, \sigma_k\rangle &= \langle \nabla f(v_{k+1}) - \nabla f(v_k), v_{k+1} - v_k\rangle \\ &= s_k \langle \nabla f(v_{k+1}) - \nabla f(v_k), d_k\rangle \\ &= s_k \langle \nabla f(v_{k+1}), d_k\rangle - s_k \langle \nabla f(v_k), d_k\rangle \\ &\geq s_k(\varepsilon_2 - 1)\langle \nabla f(v_k), d_k\rangle \ (\text{ 2nd Wolfe rule}), \\ &> 0, \end{aligned}$$

since $\varepsilon_2 < 1$ and $\langle \nabla f(v_k), d_k\rangle < 0$.

Exercise 9.3

We want to solve $-B_{k+1}\nabla f(x_k)$ if B_{k+1} is defined by the DFP recurence relationship

$$B_{k+1} = \left(I - \rho_k \sigma_k y_k^T\right) B_k \left(I - \rho_k y_k \sigma_k^T\right) + \rho_k \sigma_k \sigma_k^T. \text{ with } \rho_k = \frac{1}{\langle y_k, \sigma_k\rangle}$$

1. Let $q_k = -\nabla f(x_k)$ and

$$q_i = \left(I - \rho_i y_i \sigma_i^T\right) q_{i+1}.$$

Denote now $z_i = B_i q_i$, show that

$$z_{i+1} = z_i + (\alpha_i - \beta_i)\,\sigma_i, \text{ with } \alpha_i = \rho_i \langle \sigma_i, q_{i+1}\rangle \text{ and } \beta_i = \rho_i \langle y_i, z_i\rangle$$

2. If $L = (\sigma_k, y_k, \rho_k)_k$ is saved in a sequence. Justify the following algorithm

(a) $q = -\nabla f(x_k)$ and create an empty list called L_α
(b) For (σ, y, ρ) in reversed order of L :
  i. Compute $\alpha = \rho\langle\sigma, q\rangle$ and append α to L_α
  ii. Set $q = q - \alpha y$
(c) Reverse the list of L_α.
(d) Set $q = B_0 q$.
(e) For $(\sigma, y, \rho), \alpha$ in (L, L_α) :
  i. Compute $\beta = \rho\langle y, q\rangle$.
  ii. Set $q = q + (\alpha - \beta)\sigma$

♣ 9.9 Exercises

Exercise 9.4: Gradient method and Newton method on a simple example.

Consider the problem:

$$\min_{x\in\mathbb{R}} f(x) := \sqrt{1+x^2}. \qquad (\mathcal{P})$$

The function f is derivable on $\mathbb{R}$ and its first order derivative:

$$\forall x \in \mathbb{R},\ f'(x) = \frac{x}{\sqrt{1+x^2}}, \qquad f''(x) = \frac{1}{(1+x^2)^{\frac{3}{2}}}.$$

1. Show that $x^* = 0$ is the unique global minimizer of $(\mathcal{P})$.
2. **Gradient descent: choosing the right step.**
   (a) Write down one iteration of the gradient descent for the problem $(\mathcal{P})$. We denote by s_k the steps and by $(x_k)_{k\in\mathbb{N}}$ the sequence points chosen by the algorithm.
   (b) We test 3 different choices of step:

$$s_k = sf(x_k), \qquad \text{avec: } s \in \{1, 2, \frac{1}{2}\}.$$

   For each different value of s_k, compute x_k in terms of x_0 and study the convergence of the sequence $(x_k)_{k\in\mathbb{N}}$. Conclude.
3. **Local convergence of Newton method**
   (a) Write down one iteration of the Newton method for the problem $(\mathcal{P})$.We denote by $(x_k)_{k\in\mathbb{N}}$ the sequence points chosen by the algorithm.
   (b) Compute x_k as a function and x_0 and study the convergence of $(x_k)_{k\in\mathbb{N}}$. Conclude.

Exercise 9.5: Optimal step gradient algorithm

Consider the problem:

$$\min_{(x,y)\in\mathbb{R}^2} f(x,y) := \frac{1}{2}x^2 + \frac{9}{2}y^2. \qquad (\mathcal{P})$$

1. Is there a solution to $(\mathcal{P})$? is it unique ?
2. Write down one iteration of the gradient descent for the problem $(\mathcal{P})$. We denote by s_k the steps and by $(x_k)_{k\in\mathbb{N}}$ the sequence points chosen by the algorithm.
3. Denote $X = (x,y)^\top$. Write f as $f(X) = \frac{1}{2}X^\top AX$ with some matrix A. Compute $\nabla f(X_k)$.
4. Compute $\varphi(s) = f(X_k - s\nabla f(X_k))$ and solve the problem

$$\min_{s>0} \varphi(s).$$

5. We choose $X_0 = (9,1)^\top$ as initial point.
   (a) Prove by induction that the optimal step verifies $\forall k \in \mathbb{N},\ s_k = \frac{1}{5}$, and that the sequence $X_k = (x_k, y_k)^\top$ of points is given by:

$$X_k = \left(\frac{4}{5}\right)^k \begin{pmatrix} 9 \\ (-1)^k \end{pmatrix}.$$

   (b) Give $X^\star$, the solution of $(\mathcal{P})$. Compute the error $\|X_k - X^\star\|_2$. What type of convergence is it ?

Exercise 9.6: Least square method.

Let (x_i, y_i), $i = 1, \ldots, n$ be a cloud of points. We look for a relationship between the values x_i and y_i. In order to illustrate this exercice, we introduce two datasets:

Dataset A. We are given 5 fossile skeletons of different size of an extinct animal. We claim that there must exist a linear relationship between the length of two of their bones, namely the femur and the humerus. The data for the 5 different skeletons are

femur (x_i)	38	56	59	64	74
humerus (y_i)	41	63	70	72	84

Dataset B. The second dataset consists of:

x_i	0	1	2	3	4	5	6
y_i	−5.3	0.4	2.6	4.3	3	0.5	−5.4

1. We first suppose that there exists an affine relationship $y = ax + b$ between the data.
   (a) Formulate the problem of finding $Z = (a,b)$ as an optimization

problem of the following form

$$\min_{Z\in\mathbb{R}^2} \frac{1}{2}\|BZ+d\|^2. \qquad (\mathcal{P})$$

Make explicit the matrix B and the vector d

(b) Show that problem $(\mathcal{P})$ admits a unique global minimum point $\bar{Z}=(\bar{a},\bar{b})$ on $\mathbb{R}^2$ and show that $\bar{Z}$ is the solution of a linear problem.

(c) Prepare a python program that solves the linear regression problem $(\mathcal{P})$.

2. Given the results for the dataset B, we rather think that the relationship between x_i and y_i, $i=1,\dots,7$ is given by:

$$y=ax^2+bx+c$$

(a) Reformulate the above problem as a least square linear problem.

(b) Show that $Z=(a,b,c)$ is the solution of a linear system.

(c) Prepare a python program that solves the linear regression problem for the datasets A and B. Comment.

Exercise 9.7: Non linear least square.

Consider an RLC circuit in steady state (i.e. when equilibrium has been achieved) and in forced oscillation (i.e. when a sinusoidal voltage is imposed). We then know (from physics and by solving a second-order linear differential equation with constant coefficients) that the voltage across the capacitor is of the form:

$$U(t)=U_{max}\cos(\omega t+\phi)\,, \qquad (9.6)$$

where U_{max}, ω and ϕ depend of the circuit characteristics (resistance, inductance, capacitor, imposed voltage). The voltage across the capacitor is experimentally measured. Denote U_n the voltage measured at time t_n for n varying from 1 to N. From these N measurements, we aim at determining the characteristics of the RLC circuit, i.e. the quantities U_{max}, ω and ϕ. So we'll be looking for values of U_{max}, ω and ϕ such that $U(t_n)$ is as close as possible to U_n for all n.

1. Explain why the problem described above translates mathematically into

$$min\left\{ f(x_1,x_2,x_3)=\frac{1}{2}\sum_{n=1}^{N}\left(x_1\cos(x_2t_n+x_3)-U_n\right)^2\,,\,(x_1,x_2,x_3)\in\mathbb{R}^3\right\}$$

2. Translate this minimization problem into a problem of the type $g(x_1,x_2,x_3)=0$. Give the function g explicitly.
3. Write down Newton's algorithm for solving the problem described in b). What difficulties will you encounter when implementing this program? What alternative do you propose? You'll make explicit every linear systems to which you'll relate.

CHAPTER 10

Constrained smooth optimization

♠ 10.1 Projected gradient

Proposition 10.1.1 — Projection on a convex set. Suppose that K is a closed convex set. For every v, solve $\min_{y\in K} \|y-v\|^2$. There is a unique solution denoted $p_K(v)$. This solution solves $\langle p_K(v)-v, y-p_K(v)\rangle \geq 0 \forall y \in K$

Proof

Existence of the solution follows from the fact that $y \mapsto \|y-v\|^2$ is infinite at infinite. For uniqueness, suppose x_1 and x_2 are solutions. By convexity of K, $x_m = \frac{x_1+x_2}{2} \in K$ and

$$\|x_m - v\|^2 < \frac{1}{2}\|x_1 - v\|^2 + \frac{1}{2}\|x_2 - v\|^2 = \min_{y\in K} \|y-v\|^2.$$

Which is impossible. $\forall y \in K$, $d = y - p_K(v)$ is an admissible direction and we must have $\langle \nabla J(p_K(v)), d\rangle \geq 0$.

Definition 10.1.1 — Projected gradient algorithm. If K is a closed convex set. At each iteration k, choose a step s_k and perform the iteration

$$v_{k+1} = p_K(v_k - s_k \nabla J(v_k))$$

Then if $d_k = v_{k+1} - v_k \neq 0$, it is a direction of descent and for every $0 \leq \alpha \leq 1$, we have $v_k + \alpha d_k \in K$.

Proof

v_k and $v_{k+1} = v_k + d_k$ are in K which is convex. So that $v_k + \alpha d_k \in K$ for every $0 \leq \alpha \leq 1$. Take $\langle p_K(v)-v, y-p_K(v)\rangle \geq 0$ with $v = v_k - s_k \nabla J(v_k)$ and

$y = v_k$, we have

$$\langle v_{k+1} - (v_k - s_k \nabla J(v_k)), v_k - v_{k+1} \rangle \geq 0 \Longrightarrow \langle d_k + s_k \nabla J(v_k), -d_k \rangle \geq 0$$

$$\langle \nabla J(v_k), d_k \rangle \leq -\frac{1}{s_k} \|d_k\|^2 \leq 0$$

Proposition 10.1.2 If J has a L-Lipschitz gradient and is bounded from below and K is convex, then the fixed step projected gradient algorithm with $0 < s < \frac{2}{L}$ converges in the sense $\sum \|v_{k+1} - v_k\|^2 < +\infty$.

Proof

Start with $J(v_{k+1}) \leq J(v_k) + \langle \nabla J(v_k), v_{k+1} - v_k \rangle + \frac{L}{2} \|v_{k+1} - v_k\|^2$ and remember that if $d_k = v_{k+1} - v_k$, we have

$$\langle \nabla J(v_k), d_k \rangle \leq -\frac{1}{s} \|d_k\|^2 \leq 0.$$

Hence :$J(v_{k+1}) \leq J(v_k) + (\frac{L}{2} - \frac{1}{s}) \|d_k\|^2$ and proceed as usual.

♠ 10.2 Newtonian methods

♠ 10.2.1 Introduction to SQP

Definition 10.2.1 — Definition of SQP. Suppose that $K = \{v \text{ s.t. } h(v) = 0\}$ introduce $\mathcal{L}(v, \mu) = J(v) + \mu \cdot h(v)$. The SQP method is to apply a Newton method to solve KKT, which is

$$\nabla \mathcal{L}(v, \mu) = \begin{pmatrix} \nabla J(v) + \sum_j \mu_j \nabla h_j(v) \\ h(v) \end{pmatrix} = 0.$$

Proposition 10.2.1 If H_k is an approximation of $HJ(v_k)$, finding μ^{k+1} and $d_k = v_{k+1} - v_k$ amounts to solve the equations :

$$\begin{cases} H_k d_k + \sum_j \mu_j^{k+1} \nabla h_j(v_k) = \nabla J(v_k) \\ h_j(v_k) + \langle \nabla h_j(v_k), d_k \rangle = 0 \end{cases}$$

Proof. Newton algorihtm is $v_{k+1} = v_k + d_k$, where

$$\begin{pmatrix} H_k & J_{v_k} h^T \\ J_{v_k} h & 0 \end{pmatrix} \begin{pmatrix} d_k \\ \mu^{k+1} - \mu^k \end{pmatrix} = - \begin{pmatrix} \nabla J(v_k) + \sum_j \mu_j^k \nabla h_j(v_k) \\ h(v_k) \end{pmatrix}$$

$$\begin{cases} H_k d_k + \sum_j \mu_j^{k+1} \nabla h_j(v_k) = -\nabla J(v_k) \\ h_j(v_k) + \langle \nabla h_j(v_k), d \rangle = 0 \end{cases}$$

■

Proposition 10.2.2 — Expression of SQP algorithm. The SQP algorithm can be reformulated as follows: let H_k be an approximation of the Hessian of J at point v_k. Assume that H_k is positive definite and that the family $(\nabla h_j(v_k))_j$ is linearly independent. Find d_k and the corresponding Lagrange multiplier $\mu^\star$ that verifies

$$\min_{h_j(v_k)+\langle \nabla h_j(v_k), d\rangle = 0} \langle \nabla J(v_k), d\rangle + \frac{1}{2}\langle H_k d, d\rangle \tag{10.1}$$

Update $v_{k+1} = v_k + d_k$ and $\mu^{k+1} = \mu^\star$.

Proof. Denote $f : d \mapsto \langle \nabla J(v_k), d\rangle + \frac{1}{2}\langle H_k d, d\rangle$ and the function f is continuous, coercive and we minimize on a closed set. Hence there exists a global minimizer. Uniqueness of the minimizer stems from the fact that f is striclty convex and the constraint set is convex. The constraints are qualified by hypothesis and KKT equations are

$$\begin{cases} H_k d_k + \nabla J(v_k) + \sum_j \mu_j^\star \nabla h_j(v_k) = 0 \\ h_j(v_k) + \langle \nabla h_j(v_k), d_k\rangle = 0 \end{cases}$$

which is exactly the equation of Proposition 10.2.1. ■

Definition 10.2.2 If $K = \{v \text{ s.t. } h(v) = 0, g(v) \leq 0\}$, then the SQP method amounts to find $(d^\star, \mu^\star, \lambda^\star)$ the solution of

$$\min_{\begin{cases} h_j(v_k) + \langle \nabla h_j(v_k), d\rangle = 0 \\ g_i(v_k) + \langle \nabla g_i(v_k), d\rangle \leq 0 \end{cases}} \langle \nabla J(v_k), d\rangle + \frac{1}{2}\langle H_k d, d\rangle$$

and to set $v_{k+1} = v_k + d^\star$, $\mu^{k+1} = \mu^\star$ and $\lambda^{k+1} = \lambda^\star$

♠ 10.3 Penalization

Definition 10.3.1 — Definition of penalization. Replace the problem $\min_{v\in K} J(v)$ by $\min_{v\in V} J_\varepsilon(v)$, where

$$J_\varepsilon(v) = J(v) + \varepsilon\pi(v),$$

with a regular function $\pi \geq 0$ and one of the following choices

- **Inner penalization :** π is smooth, $\pi = +\infty$ outside X and $\varepsilon \to 0$
- **Outer penalization :** π is smooth, $\pi = 0$ inside X and $\varepsilon \to +\infty$

Exercise 10.1

If $K = \{v, g(v) \leq 0\}$, consider the following standard penalizations

- **Inner penalization :** $\pi(v) = \sum_i \log(-g_i(v))$
- **Outer penalization :** $\pi(v) = \sum_i \max(g_i(v), 0)^2$

If J is coercive and g is continuous, show that the solutions of $\min_{v\in V} J_\varepsilon(v)$ converge up to a subsequence towards the solutions of $\min_{v\in K} J(v)$.

Proposition 10.3.1 — Outer penalization convergence. Suppose that J is continuous and infinite at infinite on V and g is continuous. Let $J_\varepsilon(v) = J(v) + \varepsilon\sum_i \max(g_i(v),0)^2$ and v_ε be a global minimizer of J_ε. Then there exists $v^\star$, a global minimizer of J on K such that v_ε converges up to a subsequence to $v^\star$ as $\varepsilon \to +\infty$.

Proof. The existence of a x, a global minimizer of J on K and v_ε is standard. We have

$$J(v_\varepsilon) \leq J_\varepsilon(v_\varepsilon) \leq J(x)$$

Hence v_ε is bounded (because J is infinite at infinite on V), hence converges up to a subsequence to some $v^\star$. But

$$0 \leq \pi(v_\varepsilon) = \varepsilon^{-1}(J_\varepsilon(v_\varepsilon) - J(v_\varepsilon))$$

when $\varepsilon \to +\infty$, $\pi(v^\star) = 0$, hence $v^\star \in K$ and then $J(v^\star) \leq J(x)$ says that $v^\star$ is a global minimizer. ■

Proposition 10.3.2 If in addition J and g are C^1 and that the family $(\nabla g_i(v^\star))_{g_i \text{ active}}$ is linearly independant, then denote λ the Lagrange multiplier associated $v^\star$, a minimizer of $\min_K J$, we have

$$\lim_{\varepsilon\to+\infty} 2\varepsilon \max(g_i(v_\varepsilon),0) = \lambda_i$$

Proof. If $g_i(v^\star) < 0$ then $g_i(v_\varepsilon) < 0$ for ε large. Hence $2\varepsilon\max(g_i(v_\varepsilon),0) = 0$ for large ε and everything is proven. We have

$$0 = \nabla J_\varepsilon(v_\varepsilon) = \nabla J(v_\varepsilon) + 2\varepsilon\sum_i \max(g_i(v_\varepsilon),0)\nabla g_i(v_\varepsilon)$$

Let $\lambda_i^\varepsilon = 2\varepsilon\max(g_i(v_\varepsilon),0)$ and I the set of active constraints at $v^\star$, we have that $\nabla g_i(v_\varepsilon))_{i\in I}$ is linearly independante for large ε and that $\nabla J(v^\star) + \sum_i \lambda_i \nabla g_i(v^\star)$. Uniqueness of λ allows us to conclude that λ^ε goes towards λ. ■

Proposition 10.3.3 — Inner penalization convergence. Suppose that J is continuous and infinite at infinite on K and g is continuous. Let $J_\varepsilon(v) = J(v) + \varepsilon\sum_i |\log(-g_i(v))|$ and v_ε be a global minimizer of J_ε. Then there exists $v^\star$, a global minimizer of J on K such that v_ε converges up to a subsequence to $v^\star$ as $\varepsilon \to 0$.

Proof. The existence of a x, a global minimizer of J on K and v_ε is standard. We have $v_\varepsilon \in K$

$$J(x) \leq J(v_\varepsilon) \leq J_\varepsilon(v_\varepsilon) \leq J_\varepsilon(x) \leq J_1(x)$$

Hence v_ε is bounded (because J_1 is infinite at infinite on V), hence converges up to a subsequence to some $v^\star \in K$. We let ε goes to 0, we have

$$J(x) \leq J(v^\star) \leq J(x)$$

and $v^\star$ is a global minimizer. ■