AN INTRODUCTION TO OPTIMIZATION

LUCA NENNA

Contents

These lecture notes are based on the ones by Rémy Rodiac.

1. Introduction

To optimize is to make something as good as possible. This is the idea of finding the "best possible way". Mathematically an optimization problem can take the following form: let X be a set and let $F : X \to \mathbb{R}$ be a function (usually called *cost function*). We are interested in the following quantity

$$m := \inf_{x \in X} F(x), \tag{1.1}$$

and we want to know

(1) if the infimum is achieved (i.e., if there exists a minimizer of F in X),
(2) when the infimum is achieved, we want to know if the minimizer is unique.
(3) We want to give the most accurate description of the solution.
(4) We want to compute the solution with the help of algorithms.

> **Remark.** *Minimizing or maximizing are two equivalent problems. Indeed* $\inf_{x\in X} F(x) = -\sup_{x\in X}(-F(x))$.

Here are some examples of optimization problem.

EXAMPLE 1: *The shortest path and dynamic programming*

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and $\omega : \mathcal{E} \to \mathbb{R}^+$ be a nonnegative weight function. We recall that $\mathcal{V}$ denotes the set of vertices of the graph whereas $\mathcal{E}$ stands for the set of edges. We define a *path* as a sequence of vertices $p = (v_1, \cdots, v_n) \in \mathcal{V}^n$ such that two successive

Date: 2022-2023.

vertices v_i and v_{i+1} in p are connected by an edge in the graph, which is denoted $[v_1, v_{i+1}]$. The length of a path is defined by

$$L(p) = \sum_{i=1}^{n-1} \omega([v_i, v_{i+1}]).$$

A shortest path from $v \in \mathcal{V}$ to $w \in \mathcal{V}$ is a path from v to w that minimizes the length L. Finding shortest paths for a given graph and a given weight function can have the following applications:

- *Road networks:* vertices represent road intersections, edges are road segments between those nodes and the weight can be the length of the road, the time necessary to traverse it or the cost (in highways).
- *Social networks:* people are vertices and edges are for instance friendship relations. Find the shortest path (i.e., degree of separation) between two persons.

The (single-source) shortest path problem consists in finding the shortest paths connecting a given vertex to all other vertices. Loosely speaking, dynamic programming consists in enumerating and testing cleverly the possibilities. In this problem, the key lies in the following observation: if p is a shortest path, then any subpath of p is also a shortest path. Hence it is enough to consider paths that are extension of shortest paths. In other words:

1) Start from the source vertex and give to all neighbouring vertices the weight of the edge connecting them to the source.
2) Visit each of these neighbouring vertices and at each of them do the following: compute its neighbouring points and attribute to each of them the weight of the connecting edge plus the value of the vertex.
3) Iterate the process.

This is a rough description of the *Dijkstra algorithm.* For more information on this algorithm and for the description of another algorithm called Fast Marching Algorithm we refer to `http://www.numerical-tours.com/matlab/fastmarching_0_implementing/` (webpage of Gabriel Peyré and his Numerical Tours).

EXAMPLE 2: *Dido's problem and the calculus of variations*

Dido's problem: how to enclose a surface of maximal area inside a straight line (the north African coast line) and a rope (made by slicing the hide of a bull into very small strips and tying them together)? Mathematically, the straight line is a segment of extremities a and b to be determined and we assume that the rope of fixed length $L > 0$ can be described as the graph of a function $u : [a, b] \to \mathbb{R}^+$ (this actually excludes some configurations, which ones?). If moreover u is C^1 in (a, b), then the length constraint rewrites

$$L = \int_a^b \sqrt{1 + u'(x)^2}dx, \tag{1.2}$$

and the area to maximize is then $A = \int_a^b u(x)dx$. The solution is that Dido should enclose a half circle (a and b being adapted to the length L). We will see in this course tools to prove this result.

An image processing problem: the Rudin-Osher-Fatemi denoising model: Given a noisy image $u_0 : \Omega \to \mathbb{R}$, ($\Omega \subset \mathbb{R}^2$ being a bounded open set), given $\lambda > 0$, find u minimizing

$$\int_\Omega |\nabla u(x)|dx + \lambda \|u - u_0\|^2_{L^2(\Omega)}.$$

The first term is a *regularization term*: it penalizes the discontinuity of u, the second term is the *data attachment term*: it ensures that u is close enough to the initial image u_0.

EXAMPLE 3: *The optimal assignment problem and linear programming*

Three people x_1, x_2, x_3 are respectively in Paris, Toulouse and Marseille and they need to collect products in Lyon (y_1), Strasbourg (y_2) and Grenoble (y_3). They want to minimize the total cost of the trips knowing that

From / To	Lyon	Strasbourg	Grenoble
Paris	50	80	70
Toulouse	80	120	70
Marseille	40	80	50

Who should go where?

Let us rewrite the problem, if c_{ij} denotes the cost when x_i goes to the town y_i, then the tabular above is the matrix $(c_{ij})_{ij}$. Let $A = (a_{ij})_{ij}$ be defined as $a_{ij} = 1$ if x_i goes to y_j and 0 otherwise. We thus want to find A minimizing

$$\sum_{i,j=1}^{3} a_{ij} c_{ij} = \mathrm{tr}(AC^T)$$

under the constraints

i) $a_{ij} \in \{0, 1\}$
ii) $\sum_j a_{ij} = 1$ (person i goes to exactly one town)
iii) $\sum_i a_{ij} = 1$ (town j is reached by exactly one person.

It is not obvious but the constraint $a_{ij} \in \{0, 1\}$ can be replaced by

$$a_{ij} \in [0, 1],$$

without changing the minimizers. Then we are reduced to minimize a *linear cost* under *linear constraints*. In this simple case, it is possible to test all possibilities (3!), but in general, the number of configurations to test would be $n!$ (for n people going to n cities), which is not numerically possible. There exists an algorithm called the *Hungarian algorithm* which allows to solve this problem in $O(n^3)$. We will not see linear optimization in this course, we refer to [5] chap.10.

Bibliography: As a short bibliography we refer to [4], [5], [1], [2] and also to [3] for the functional analysis results used in this course.

2. Existence of minimizers

Throughout this course, we shall set

- a Banach space $(V, \|\cdot\|)$,
- a non-empty subset $A \subset V$,
- a cost function $J : A \to \mathbb{R}$.

The optimization problem we consider is: does there exists $x_* \in A$ such that

$$J(x_*) = \inf_{x \in A} J(x)?$$

If it is the case we write $\inf_{x \in A} J(x) = \min_{x \in A} J(x) = J(x_*)$. It is important to realize that existence of minimizers is not always true. Roughly speaking, two problems can occur:

1) the minimization set A is not compact (example: $J(x) = e^{-x}$, $A = V = \mathbb{R}^+$);

2) the cost function J is not "continuous enough" (example: $J(x) = -x$, if $x \in [-1, 0[$ and $J(x) = 2 - x$, if $x \in [0, 1]$, here $A = [-1, 1] \subset \mathbb{R}$).

Existence results will be given, with hypothesis ruling out these problems (some kind of compactness and lower semi-continuity). These results can be viewed as generalizations of the following theorem: *Let $K \subset \mathbb{R}^n$ be a compact set and $f : K \to \mathbb{R}$ be a continuous function, then f is bounded and the infimum and supremum of f in K are achieved.* For the uniqueness question, we will see that uniqueness holds essentially under strict convexity assumptions. We start by recalling some definitions.

2.1. Some definitions and notations.

Definition 2.1. (Global, local minimizer) An element $x_* \in A$ is a global minimizer of J on A if and only if (iff)

$$\forall x \in A, \quad J(x_*) \leq J(x).$$

An element $x_* \in A$ is a local minimizer of J on A if and only if

$$\exists \delta > 0,\ \forall x \in A,\ \|x - x_*\| \leq \delta \Rightarrow J(x_*) \leq J(x).$$

Definition 2.2. (Minimizing sequences). A minimizing sequence of J in A is a sequence $(x_n)_n \subset A$ such that $J(x_n) \to \inf_A J$ when $n \to +\infty$.

Remark. *From the definition of the infimum a minimizing sequence always exists.*

2.2. Lower semi-continuity.

We give definitions in the general framework of a metric space (X, d).

Definition 2.3 (Lower semi-continuity (l.s.c)). Let (X, d) be a metric space and $f : X \to \mathbb{R} \cup +\infty$, then f is said to be lower semi-continuous (l.s.c) at $a \in X$ if for every $\epsilon > 0$ there exists an $r > 0$ such that $f(x) \geq f(a) - \epsilon$ for all $x \in B(a, r)$.

In other words, comparing with the definition of continuity at a, the difference is that lower semi-continuity only requires one inequality $f(x) \geq f(a) - \epsilon$ while continuity requires both $f(x) \geq f(a) - \epsilon$ and $f(x) \leq f(a) + \epsilon$. Lower semi-continuity can be seen as continuity but only when coming to a from below. Lower semi-continuity can be characterized in terms of the epigraph.

Definition 2.4 (Epigraph). Let (X, d) be a metric space and $f : X \to \mathbb{R} \cup \{+\infty\}$, then the epigraph of f is the set of points lying above its graph:

$$\operatorname{epi}(f) = \{(x, \lambda) \in V \times \mathbb{R}\ ;\ \lambda \geq f(x)\}.$$

Proposition 2.5. *Let $f : X \to \mathbb{R} \cup \{+\infty\}$, then f is lower semi-continuous iff* $\operatorname{epi}(f)$ *is closed in $X \times \mathbb{R}$.*

Exercise 1. *Prove the above statement.*

Definition 2.6. (Sequential lower semi-continuity) Let (X,d) be a metric space and $f : X \to \mathbb{R} \cup \{+\infty\}$, then f is said to be lower semi-continuous at $a \in X$ if for all sequence $(x_n)_n$ tending to a in X,

$$\liminf_{n\to+\infty} f(x_n) \geq f(a).$$

Remark (Continuity vs. sequential continuity)**.** *Recall that, while (semi) continuity always implies sequential (semi) continuity, the converse implication is not true in general. However the equivalence is true in metric spaces. Moreover, in all the results of existence of minimizers stated in these notes, only the sequential lower semi-continuity is needed.*

2.3. **The direct method in the calculus of variations.** Let (X,d) be a metric space. The direct method in the calculus of variations is a method to prove existence of minimizers which consists in the following steps.

1) Take a minimizing sequence $(x_n)_n \subset A$ and show that it admits a subsequence $(x_{n_k})_k$ converging, in some sens to be defined, to some $x_* \in A$. This is a compactness issue.
2) Show that J is sequentially lower semi-continuous.
3) In this case,

$$\inf\{J(x); x \in A\} = \lim_{n\to\infty} J(x_n) = \lim_{k\to\infty} J(x_{n_k}) \geq J(x_*).$$

In the following we will apply this method to prove more specific existence results. We will make a distinction between finite dimensional problems and infinite dimensional problems. The reason is that in finite dimension, compact sets are easy to characterize, they are closed bounded sets, whereas this property is not true in infinite dimension. We recall

Definition 2.7 (Coercivity)**.** Let V be a normed vector space. A function $f : A \subset V \to \mathbb{R}$ is coercive (or infinite at infinity) iff

$$\lim_{\substack{\|x\|\to+\infty \\ x\in A}} f(x) = +\infty.$$

2.4. **Existence in finite dimension.**

Theorem 2.8 (Existence in finite dimension)**.** *Let V be a normed vector space of finite dimension. Assume that $J : A \to \mathbb{R}$ is l.s.c and coercive, and assume that A is non empty and closed, then there exists at least one minimizer of J in A.*

The proof is a direct application to the direct method of the previous paragraph (write the details to make sure you understood!)

2.5. **The case of a quadratic functional.** This is a very important particular case. Let $(V, \langle,\cdot,\cdot,\rangle)$ be a **Hilbert space** and $J : V \to \mathbb{R}$ be defined as

$$J(x) = \frac{1}{2}a(x,x) - b(x), \tag{2.1}$$

where $a : V \times V \to \mathbb{R}$ is a symmetric continuous bilinear form and $b : V \to \mathbb{R}$ is a continuous linear form. When J has this particular form, existence and uniqueness of a minimizer can be proved by Hilbertian methods, it follows from the projection theorem and Riesz' representation theorem (in particular we do not need the direct method in this case). Existence and uniqueness hold under the assumption that the bilinear form a is *elliptic*.

Definition 2.9. A bilinear form $a : V \times V \to \mathbb{R}$ is elliptic iff there exists $\alpha > 0$ such that for all $x \in V$,

$$a(x, x) \geq \alpha \|x\|^2.$$

Theorem 2.10. *Let $J : V \to \mathbb{R}$ be a quadratic functional defined on the Hilbert space V as in (2.1). Assume that the bilinear form a is elliptic and that A is a convex closed subset of V. Then there exists a unique solution x_* to the minimization problem*

$$J(x_*) = \min_{x \in A} J(x).$$

Moreover, x_* is the solution to the minimization problem above iff

$$a(x_*, x - x_*) \geq b(x - x_*) \text{ for all } x \in A. \tag{2.2}$$

Proof.

- As a is elliptic, it defines another scalar product on V. By Riesz theorem, there exists a unique $v \in V$ such that for all $x \in V$,
$$b(x) = a(v, x).$$
- Therefore
$$J(x) = \frac{1}{2} a(x - v, x - v) - \frac{1}{2} a(v, v),$$
$v \in V$ is fixed and thus it is equivalent to minimize $a(x - v, x - v) = \|x - v\|_a^2$ for $x \in A$, where $\| \cdot \|_a$ denotes the norm associated with the scalar product a. This exactly amounts to look for the projection of v onto A.
- As A is a closed convex set, the projection theorem ensures that there exists a unique solution $x_* \in A$ characterized by
$$a(v - x_*, x - x_*) \leq 0.$$

□

Remark. *In the case $A = V$, characterization (2.2) simply rewrites: for all $x \in V$,*

$$a(x_*, x) = b(x).$$

Moreover, characterization (2.2) is an optimality condition of first order. We will come back on that in the next chapter.

EXAMPLE 1 (Least square approximation): Let $A \in M_{n,p}(\mathbb{R})$ and $B \in M_{n,1}(\mathbb{R})$ with $n > p$. The linear system $Ax = B$ is generally overdetermined and thus may not have a solution. The idea is then to consider a generalized solution, for example we can consider the solution in the sense of *least square approximation*. This is defined as the minimizer $x_* \in \mathbb{R}^p$ of

$$J(x) = \frac{1}{2}\|Ax - B\|^2 = \frac{1}{2}\langle A^T Ax, x\rangle - \langle A^T B, x\rangle + \frac{1}{2}\|B\|^2.$$

This is a finite dimensional problem. The matrix A^TA is symmetric and it is positive definite if A has rank p. In this case A^TA defines an elliptic bilinear form and we can apply Theorem 2.10 to obtain existence and uniqueness of the solution. Furthermore, characterization (2.2) gives what is called the *normal equation*

$$A^TAx_* = A^TB.$$

The least square approximation can be applied in the case of a polynomial fitting (of order 3 for instance). Given n points $\{(x_i, y_i)\}_{i=1,\ldots,n} \subset \mathbb{R}^2$ we look for the polynomial curve of order 3 which fits in the best way those points. This amounts to look for the coefficients $\alpha, \beta, \gamma, \delta$ which minimize $\sum_{i=1}^n |y_i - \alpha + \beta x_i + \gamma x_i^2 + \delta x_i^3|^2$. This can be written in the form of a least square approximation problem as before with

$$A = \begin{pmatrix} 1 & x_1 & (x_1)^2 & (x_1)^3 \\ \vdots & & & \vdots \\ 1 & x_n & (x_n)^2 & (x_n)^3 \end{pmatrix}, \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and } x = \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix}.$$

EXAMPLE 2: (Variational formulation of elliptic problem) Let $\Omega \subset \mathbb{R}^n$ be a smooth[1] bounded open set and $f \in L^2(\Omega)$. Let $J : H_0^1(\Omega) \to \mathbb{R}$ be defined as

$$J(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega fu.$$

Then there exists a unique u minimizing J in H_0^1 and moreover u is solution of the variational formulation in H_0^1 of the Poisson equation with Dirichlet boundary conditions

$$\begin{cases} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{cases}$$

Indeed, let a and b be the applications defined as

$$\begin{array}{rlll} a & : H_0^1(\Omega) & \times H_0^1(\Omega) \to \mathbb{R} \\ & (u,v) & \mapsto \frac{1}{2}\int_\Omega \langle \nabla u, \nabla v\rangle \end{array} \quad \text{and} \quad \begin{array}{rlll} b & : H_0^1(\Omega) & \to & \mathbb{R} \\ & u & \mapsto & \int_\Omega fu. \end{array}$$

- The linear form b is continuous: for every $u \in H_0^1(\Omega)$, $|b(u)| \leq \|f\|_{L^2}\|u\|_{L^2} \leq \|f\|_{L^2}\|u\|_{H_0^1}$.
- The symmetric bilinear form a is continuous: for every $u, v \in H_0^1(\Omega)$, $|a(u,v)| \leq \|\nabla u\|_{L^2}\|\nabla v\|_{L^2} \leq \|u\|_{H_0^1}\|v\|_{H_0^1}$.
- a is elliptic: thanks to Poincaré's inequality, there exists a constant $C > 0$ depending only on Ω, such that $\|u\|_{H_0^1} \leq C\|\nabla u\|_{L^2}$. Therefore, for $u \in H_0^1(\Omega)$,

$$a(u,u) \geq \frac{1}{2C^2}\|u\|_{H_0^1}^2.$$

We can then apply Theorem 2.10 to obtain a unique minimizer $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \langle \nabla u, \nabla v\rangle = \int_\Omega fv \quad \text{for all } v \in H_0^1(\Omega).$$

This is the variational formulation of the Poisson equation. The boundary condition $u = 0$ on $\partial\Omega$ is encoded in the choice of the space $H_0^1(\Omega)$.

In infinite dimension, when the functional is not quadratic and Hilbertian methods are not available we will rely on the direct method of calculus of variations. However being compact is stronger than being closed and unbounded. When the topology comes from a norm, Riesz Theorem (the unit ball of a Banach space is compact iff it has finite

[1] Regular enough in order to apply Poincaré's inequality, Lipschitz for instance.

dimension) even says that it is always stronger. Thus we use another topology for which some bounded closed sets are compact. Our goal is to find a topology such that *for all bounded sequences we can extract a converging subsequence.*

2.6. Weak topology and compact sets.

We recommend Chapter 3 of [3] for the results in this section. Since compactness is an essential property (for existence of minimizers, for instance) one naturally has to consider another topology than the strong one (i.e., the topology induced by the norm $\|\cdot\|$), for which there are less open sets but more compact sets. In finite dimension, the weak topology and the strong topology coincide. In infinite dimension, the weak topology is strictly smaller than the strong topology.

Proposition 2.11. *(Weakly convergent sequences) Let V be a Banach space, $V^\star$ denote its dual and $(x_n)_n \subset V$. Then,*

- *the sequence $(x_n)_n$ weakly converges to x iff for every $\zeta \in V^\star$, $\zeta(x_n) \underset{n\to+\infty}{\to} \zeta(x)$,*
- *if $(x_n)_n$ converges in norm then it weakly converges to the same limit.*

Exercise 2. *Show that the unit open ball is not weakly open while the unit sphere is not weakly closed.*

Our purpose is not to study weak topology, therefore we only state the results which we will use, we refer to [3] for more on this notion. As closed sets are complementary sets of open sets, there are also sets which are (strongly) closed but not weakly closed, however there is a case where both coincide: the case of convex sets. This is a consequence of Hahn-Banach Theorem.

Proposition 2.12. *Let $C \subset V$ be a convex set, then C is strongly closed iff it is weakly closed.*

This is a first example where the convexity plays an important role. To obtain the desired weak compactness of convex closed bounded sets we shall work with reflexive Banach spaces.

Definition 2.13. (Reflexive Banach space) A reflexive Banach space is a Banach space V isomorphic to its bi-dual $V^{\star\star}$ via the canonical injection

$$\begin{array}{rccc} \mathcal{J}: & V & \to & V^{\star\star} \\ & x & \mapsto & \left(\begin{array}{cccc} \Phi_x: & V^\star & \to & \mathbb{R} \\ & L & \mapsto & L(x) \end{array}\right). \end{array}$$

Remark: It is easy to check that the canonical injection is linear continuous and that $\|J(x)\|_{V^{\star\star}} \leq \|x\|$. Indeed,

$$\|\mathcal{J}(x)\|_{V^{\star\star}} = \sup_{\|\zeta\|_{V^\star}=1} |\Phi_x(\zeta)| = \sup_{\|\zeta\|_{V^\star}=1} |\zeta(x)| \leq \|\zeta\|_{V^\star}\|x\| \leq \|x\|.$$

Proposition 2.14. *Let V be a reflexive Banach space, then the canonical injection $\mathcal{J}$ is an isometry: for all $x \in V$,*

$$\|\mathcal{J}(x)\|_{V^{\star\star}} = \|x\|.$$

Proof. It is a consequence of Hahn-Banach Theorem. Fix $x \in V$ and consider the linear form L defined on $\mathbb{R}x \subset V$ by $L(y) = t$ if $y = t\frac{x}{\|x\|} \in \mathbb{R}x$. The linear form L is continuous and $\|L\|_{(\mathbb{R}x)'} = 1$. Indeed, $|L(y)| = |t| = \|y\|$. By Hahn-Banach theorem, there exists a continuous linear form on V, $\zeta_0 \in V^\star$, such that $\|\zeta_0\|_{V^\star} = \|L\|_{(\mathbb{R}x)'} = 1$ and $\zeta_0 = L$ in $\mathbb{R}x$. In particular, $\zeta_0(x) = L(x) = \|x\|$. We can thus conclude that

$$\|\mathcal{J}(x)\|_{V^{\star\star}} = \sup_{\|\zeta\|_{V^\star}=1} |\Phi_x(\zeta)| \geq |\Phi_x(\zeta_0)| = |\zeta_0(x)| = \|x\|.$$

□

EXAMPLES: The following are reflexive Banach spaces:

- finite dimensional Banach spaces,
- Hilbert spaces (in particular $L^2(\Omega), H^1(\Omega)$),
- $L^p(\Omega)$ for $1 < p < +\infty$,
- $W^{1,p}(\Omega)$ for $1 < p < +\infty$,
- every closed subspace of a reflexive Banach space.

However $L^1(\Omega), L^\infty(\Omega), C^0([0,1])$ are not reflexive.

Proposition 2.15. *Let V be a reflexive Banach space and $(x_n)_n \subset V$ a sequence weakly converging to x. Then,*

i) *the norm is weakly l.s.c.:* $\|x\| \leq \liminf_{n\to\infty} \|x_n\|$,
ii) *the sequence is bounded:* $\sup_n \|x_n\| < +\infty$.

Proof. i) It is a consequence of the strong continuity and the convexity of the norm, another proof is given in the next item. ii) It is a consequence of Banach-Steinhaus theorem and the fact that J is an isometry. Indeed, let us consider the evaluation maps $\Phi_{x_n} = \mathcal{J}(x_n)$ and $\Phi_x = \mathcal{J}(x)$. As x_n weakly converges to x, for any $\zeta \in V^\star$, $\zeta(x_n) \underset{n\to\infty}{\to} \zeta(x)$ that is, $\Phi_{x_n})(\zeta) \to \Phi_x(\zeta)$. Hence, the sequence of continuous linear forms (Φ_{x_n}) pointwise converges to Φ_x. The Banach-Steinhaus theorem implies that $\sup_n \|\Phi_{x_n}\|_{V^{\star\star}} < +\infty$ and $\|\Phi_x\|_{V^{\star\star}} \leq \liminf_{n\to\infty} \|\Phi_{x_n}\|_{V^{\star\star}}$. We conclude with the fact that $\|\Phi_{x_n}\|_{V^{\star\star}} = \|x_n\|$ and $\|\Phi_x\|_{V^{\star\star}} = \|x\|$ since J is an isometry. □

We can now state the result we were interested in, that is weak compactness of convex closed bounded sets.

Theorem 2.16. *Let V be a reflexive Banach space. Then convex bounded closed sets are weakly compact.*

2.7. **An example in infinite dimension.** Let $\Omega \subset \mathbb{R}^n$ be a regular bounded open set, $f \in L^2(\Omega)$ and define E on $H^1(\Omega)$ by

$$E(u) = \int_\Omega |\nabla u|^2 + \int_\Omega (u^2 - 1)^2 + \int_\Omega |u - f|^2.$$

We consider the problem of minimizing E in $H^1(\Omega)$. We apply the direct method in the calculus of variations to prove existence of minimizers.

- **Coercivity:** As we have
$$E(u) \geq \|\nabla u\|_{L^2}^2 + \int_\Omega (u^2-1)^2,$$
it is enough to bound from below $(u^2-1)^2$ by some term of order u^2. For instance, $(u^2-1)^2 = (u^2-2)^2 + 2u^2 - 3 \geq 2u^2 - 3$. Therefore,
$$\int_\Omega (u^2-1)^2 \geq 2\|u\|_{L^2}^2 - 3|\Omega|,$$
which leads to $E(u) \geq \|u\|_{H^1}^2 - 3|\Omega| \to +\infty$ when $\|u\|_{H^1} \to +\infty$.
- **Compactness:** Notice that E is proper (i.e., E is not constantly equal to $+\infty$, actually, in our case, $E(u) < +\infty$ for all $u \in H^1$) and let u_n be a minimizing sequence, that is $E(u_n) \underset{n\to\infty}{\to} \inf_{u\in H^1} E(u)$. From the coercivity of E, we know that $(u_n)_n$ is bounded in H^1 and thus weakly compact in H^1 from Theorem 2.16.
- **Lower semi-continuity:** Let $(u_n)_n$ be a sequence weakly converging to u in H^1. We want to prove that $E(u) \leq \liminf_{n\to\infty} E(u_n)$. Let us rewrite E as
$$E(u) = \|u\|_{H^1}^2 - 2\int_\Omega fu + \|f\|_{L^2} + \int_\Omega (u^2-1)^2.$$
As the norm is l.s.c with respect to the weak convergence (cf. Proposition 2.15) and the application $u \mapsto \int_\Omega fu$ is linear continuous in H^1 and thus H^1-weakly continuous (by definition of the weak topology), it remains to prove that
$$\int_\Omega (u^2-1)^2 \leq \liminf_{n\to\infty} \int_\Omega ((u_n)^2-1)^2.$$
Up to extraction, we can assume that
$$\liminf_{n\to\infty} \int_\Omega ((u_n)^2-1)^2 = \lim_{n\to\infty} \int_\Omega ((u_n)^2-1)^2.$$
As $(u_n)_n$ is H^1-weakly converging, it is bounded in H^1 (see Proposition 2.15). Since the injection of $H^1(\Omega)$ into $L^2(\Omega)$ is compact[2] there is a subsequence $(u_{n_k})_k$ which converges strongly in L^2 to u[3]. We can thus extract again a subsequence $(u_{n_{k_l}})_l$ which converges almost everywhere to u. By Fatou's Lemma we find
$$\int_\Omega (u^2-1)^2 \leq \liminf_{l\to\infty} \int_\Omega (u_{n_{k_l}}^2 - 1)^2 = \lim_{n\to\infty} \int_\Omega ((u_n)^2-1)^2.$$
A sum of two l.s.c. functions is l.s.c. hence E is H^1-weakly lower semi-continuous.
- **Conclusion:** We conclude with the direct method of calculus of variations. Let $(u_n)_n$ be a minimizing sequence of E in $H^1(\Omega)$, as we showed its compactness, let $(u_{n_k})_k$ be a subsequence converging to $u_* \in H^1(\Omega)$. As E is $H^1(\Omega)$-weakly lower semi-continuous, we have
$$\inf_{H^1(\Omega)} E = \lim_{n\to\infty} E(u_n) = \liminf_{k\to\infty} E(u_{n_k}) \geq E(u_*).$$

[2] This is the Rellich-Kondrachov theorem, here it is essential that Ω is compact, the injection $H^1(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n)$ is not compact.

[3] Assume that $u_{n_k} \underset{L^2}{\to} v$, as both strong convergence in L^2 and weak convergence in H^1 imply distributional convergence, $u = v$ by uniqueness of the distributional limit.

This proves that u_* is a minimizer of E in $H^1(\Omega)$.

Exercise 3. *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and consider*

$$\begin{array}{rll} L: & L^2(]0,1[) & \to \mathbb{R} \\ & u & \mapsto \int_0^1 f(u(x))dx. \end{array}$$

Show that L is weakly l.s.c. implies that f is convex.

However, in the previous example, we showed that $u \mapsto \int_\Omega (u^2-1)^2$ was weakly l.s.c. while $f(t) = (t^2-1)^2$ is not convex in $\mathbb{R}$. Is it a contradiction? Fortunately no, indeed we showed that E is H^1-weakly l.s.c and not L^2-weakly l.s.c. Actually we used a strong L^2-convergence through compact injection H^1 into L^2.

2.8. Existence in infinite dimension.

Exercise 4. *Check that $J : A \to \mathbb{R}$ is convex iff* $\operatorname{epi}(J)$ *is convex.*

Theorem 2.17. *Let V be a reflexive Banach space and let $A \subset V$ be a closed convex (non empty) set, and assume that $J : A \to \mathbb{R}$ is convex, lower semi-continuous (for the norm) and coercive, then there exists a minimizer of J in A.*

Proof.

- J convex implies that $\operatorname{epi}(J)$ is convex. Therefore: J is (strongly) lower semi-continuous iff $\operatorname{epi}(J)$ is strongly closed iff $\operatorname{epi}(J)$ is weakly closed iff J is lower semi-continuous for the weak topology.
- J is coercive so that $J(x) \geq M$ for all x such that $\|x\| \geq R$.
- Let B be the closed ball of radius R centred at 0. Then $A \cap B$ is a closed convex set. This means that it is also a weakly closed convex set and thus by Theorem 2.16, $A \cap B$ is weakly compact.
- J is weakly lower semi-continuous on a weakly compact set, thus there exists $x_* \in A \cap B$ a minimizer of J on $A \cap B$ and for all $x \in A \setminus B$, $J(x) \geq M \geq J(x_*)$.

□

We conclude this section with an example where existence of a minimizer fails.

Exercise 5. *We consider the Hilbert space*[4]

$$l^2(\mathbb{R}) = \Big\{ (x_n)_{n\in\mathbb{N}}; \sum_{n=0}^{\infty} x_n^2 < \infty \Big\},$$

endowed with the inner product $(x_n)_n \cdot (y_n)_n = \sum_{n=0}^\infty x_n y_n$. We define

$$\begin{array}{rlll} f & : l^2(\mathbb{R}) & \to & \mathbb{R} \\ & (x_n)_n & \mapsto & (\|x\|^2-1)^2 + \sum_{n=0}^\infty \frac{x_n^2}{n+1}. \end{array}$$

Check that f is coercive and lower semi-continuous for the norm. Check however that f does not admit a minimizer on $l^2(\mathbb{R})$. The problem is that f is neither convex nor weakly l.s.c.

3. Notions of differential calculus

The aim of this section is to freshen up fundamental definitions and properties in differential calculus. These will be used in the sequel.

[4] Recall that Hilbert spaces are reflexive.

3.1. **Differentiability.**

Definition 3.1. Let $(V, \|\cdot\|_V), (W, \|\cdot\|_W)$ be two normed vector space and $f : \Omega \subset V \to W$ be defined on an open set Ω. Let $x_0 \in \Omega$, f is differentiable at x_0 if and only if there exists a **continuous** linear application $L \in \mathcal{L}(V, W)$ such that

$$f(x_0 + h) \underset{h\to 0}{=} f(x_0) + L(h) + o(\|h\|_V).$$

The linear application L is denoted by $Df(x_0) \in \mathcal{L}(V, W)$ and is called the differential of f at x_0.

Remark. *In finite dimension, as all the norms are equivalent, the differentiability does not depend on the choice of norms. However, in infinite dimension, the differentiability depends on the norms on E and F. We also recall that in finite dimension a linear application is automatically continuous, which is not true in infinite dimension.*

The application f is said to be $C^1(\Omega)$ if f is differentiable at every point of Ω and the application

$$\begin{array}{rccl} Df : & \Omega & \to & (\mathcal{L}(V, W), \|\cdot\|_{\text{op}}) \\ & x_0 & \mapsto & Df(x_0) \end{array}$$

is continuous.

Definition 3.2. (Directional derivative) Let $f : \Omega \subset V \to W$ be defined in an open set Ω and let $x_0 \in \Omega$, $h \in V$. When it exists, the limit

$$\lim_{t\to 0} \frac{f(x_0 + th) - f(x_0)}{t}$$

is called the directional derivative of f along h and sometimes denoted $\partial_h f(x_0)$.

Remark. *If f is differentiable at x_0 then f admits directional derivative in any direction $h \in V$ at $x_0 \in V$ and*

$$Df(x_0).h = \lim_{t\to 0} \frac{f(x_0 + th) - f(x_0)}{t}.$$

This is the usual way to compute the differential. However the converse is not true, f can have directional derivative in all directions at x_0 and not being differentiable at x_0. For example $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \begin{cases} \frac{y^2}{x} \text{ if } x \neq 0 \\ y \text{ if } x = 0. \end{cases}$$

When f is real-valued, i.e., $W = \mathbb{R}$, which is the case of the cost function J, and V is a Hilbert space, the differential is a continuous linear form and thus can be represented by an element of V itself (Riesz representation theorem), i.e., there exists an element of V, usually denoted by $\nabla f(x_0)$ such that for all $h \in V$

$$Df(x_0).h = \langle \nabla f(x_0), h\rangle.$$

This element is called the *gradient* of f at x_0. If $V = \mathbb{R}^n$ is endowed with the usual scalar product and $(e_1, \cdots, e_n)$ is the canonical basis,

$$\nabla f(x_0) = \begin{pmatrix} \partial_1 f(x_0) \\ \partial_2 f(x_0) \\ \vdots \\ \partial_n f(x_0) \end{pmatrix}$$

where $\partial_i f(x_0) = Df(x_0).e_i$ is the directional derivative in the direction e_i, that is the i^{th} partial derivative of f at x_0.

Exercise 6. *Show that $f : (\mathbb{R}_+^*)^3 \to \mathbb{R}$ defined by $f(x_1, x_2, x_3) = x_1 \ln x_1 + x_2 \ln x_2 + x_3 \ln x_3$ is twice differentiable and compute its differential and Hessian (the matrix of the second partial derivatives).*

Exercise 7. *(An example in infinite dimension) Let $a < b$ and*

$$E_0 := \{u \in C^1([a, b], \mathbb{R}); u(a) = u(b) = 0\}$$

be the vector space normed with $\|u\|_{C^1} := \sup_{[a,b]}(|u| + |u'|)$. Let $L : \mathbb{R}^3 \to \mathbb{R}$ be C^1 and define for $u \in E_0$,

$$J(u) = \int_a^b L(x, u(x), u'(x))dx.$$

Show that J is differentiable in E_0 and compute its differential.

Solution: Let $u, h \in E_0$,

$$\begin{aligned} J(u+h) - J(u) &= \int_a^b \left[L(x, u(x) + h(x), u'(x) + h'(x)) - L(x, u(x), u'(x))\right] dx \\ &= \int_a^b \langle \nabla L(x, u(x), u'(x)), (0, h(x), h'(x))\rangle + o((0, h(x), h'(x)))dx \\ &=: l_u(h) + r_u(h), \end{aligned}$$

where for a fixed $u \in E_0$,

$$\begin{array}{rccl} l_u : & E_0 & \to & \mathbb{R} \\ & h & \mapsto & \int_a^b \langle \nabla L(x, u(x), u'(x)), (0, h(x), h'(x))\rangle dx \\ & & & = \int_a^b \partial_2 L(x, u(x), u'(x))h(x) + \partial_3 L(x, u(x), u'(x))h'(x)dx. \end{array}$$

and $r_u(h) = J(u+h) - J(u) - l_u(h)$. The application l_u is linear and continuous. Indeed, applying the triangular inequality and Cauchy-Schwartz inequality we have for all $h \in E_0$,

$$|l_u(h)| \leq (\|h\|_{C^0} + \|h'\|_{C^0}) \sup_{x \in [a,b]} |\nabla L(x, u(x), u'(x))| \leq C\|h\|_{C^1}.$$

It remains to prove that $r_u(h) = o(\|h\|_{C^1})$.

* First step: Let us prove that for all $K \subset \mathbb{R}^3$ compact set, for all $\epsilon > 0$,

$\exists \eta > 0, \ \forall X \in K, \ \forall H \in \mathbb{R}^3,$

$$\|H\| < \eta \Rightarrow |L(X+H) - L(X) - \langle \nabla L(X), H\rangle \leq \epsilon \|H\|.$$

Let $K' := \{X + H \in \mathbb{R}^3; X \in K, \ \|H\| \leq 1\}$, K' is compact. If L was C^2, we could directly apply Taylor-Lagrange expansion at order 2 in $[X, X+H]$ and use the boundedness of $\operatorname{Hess} L$ in K'. Here, as L is only assumed to be C^1, we use the uniform continuity of ∇L in K'. Let $\eta > 0$ be such that for all $X, Y \in K'$,

$$\|X - Y\| < \eta \Rightarrow \|\nabla L(X) - \nabla L(Y)\| < \epsilon.$$

For $X \in K$ and $H \in \mathbb{R}^3, \|H\| < \eta$, we apply the mean value theorem: there exists $\theta \in]0,1[$, such that

$$L(X+H) - L(X) = \langle \nabla L(X+\theta H), H \rangle.$$

As $X, X+\theta H \in K'$ and $\|X+\theta H - X\| = \theta\|H\| < \eta$, we have

$$\begin{aligned} |L(X+H) - L(X) - \langle \nabla L(X), H \rangle| &= |\langle \nabla L(X+\theta H), H \rangle - \langle \nabla L(X), H \rangle| \\ &\leq \|\nabla L(X+\theta H) - \nabla L(X)\| \|H\| \leq \epsilon \|H\|. \end{aligned}$$

* Second step: We apply the first step with $X(x) = (x, u(x), u'(x))$ and $H(x) = (0, h(x), h'(x))$ for all $x \in [a,b]$. As u and u' are continuous in $[a,b]$, there exists $R > 0$ such that for all $x \in [a,b]$, $X(x)$ is in the closed ball $K = B_R(0)$. Moreover, for all $x \in [a,b]$,

$$\|H(x)\| = \sqrt{|h(x)|^2 + |h'(x)|^2} \leq |h(x)| + |h'(x)| \leq \|h\|_{C^1}.$$

For all $\epsilon > 0$ let $\eta > 0$ be given by the first step. For $h \in E_0$ such that $\|h\|_{C^1}, \|H\| < \eta$ and consequently

$$\begin{aligned} |r_u(h)| = \Big| \int_a^b & L(x, (u+h)(x), (u'+h')(x)) - L(x, u(x), u'(x)) \\ & - \langle \nabla L(x, u(x), u'(x), (0, h(x), h'(x)) \rangle dx \Big| \\ \leq & \int_a^b |L(X(x) + H(x)) - L(X(x)) - \nabla L(X(x)), H(x) \rangle | dx \\ \leq & \int_a^b \epsilon \|H(x)\| \\ \leq & (b-a)\epsilon \|h\|_{C^1}. \end{aligned}$$

In other words, $r_u(h) = o(\|h\|_{C^1})$ and $DJ(u) = l_u$.

3.2. **Taylor formulas.** We restrict ourselves to 2nd order since it is all we will use thereafter. An application $f : \Omega \subset V \to W$ is twice differentiable at x_0 if it is differentiable in a neighbourhood $U(x_0)$ of x_0 and the application

$$\begin{array}{ccc} U(x_0) & \to & (\mathcal{L}(V,W), \|\cdot\|_{\text{op}}) \\ x & \mapsto & Df(x) \end{array}$$

is differentiable at x_0. The resulting differential

$$\begin{array}{cccc} D(Df)(x_0): & V & \to & \mathcal{L}(V,w) \\ & h & \mapsto & \begin{array}{cccc} D(Df)(x_0).h: & V & \to & W \\ & k & \mapsto & (D(Df)(x_0).h).k \end{array} \end{array}$$

is identified with the continuous bilinear application

$$\begin{array}{cccc} D^2 f(x_0): & V \times V & \to & W \\ & (h,k) \mapsto & D(Df)(x_0).(h,k) & \end{array}.$$

We now recall Taylor's formulas (of order 1 and 2).

Theorem 3.3. *(Taylor's formula) Let W be a normed vector space and $f : \Omega \subset V \to W$ be defined in an open set Ω of a vector space V and $x_0 \in \Omega$. If f is differentiable at x_0, then*

$$f(x_0+h) = f(x_0) + Df(x_0).h + o(\|h\|) \qquad \textit{Taylor-Young formla}\ ,$$

if $f \in C^1(\Omega)$ and if W is a Banach space

$$f(x_0+h) = f(x_0) + \int_0^1 [Df(x_0+th).h]dt \qquad \textit{Taylor with integral remainder}$$

If f is twice differentiable at x_0, then

$$f(x_0+h) = f(x_0)+Df(x_0).h+\frac{1}{2}D^2f(x_0).(h,h)+o(\|h\|^2) \qquad \textit{(Taylor-Young formula).}$$

If f is real valued i.e., $W=\mathbb{R}$ and twice differentiable in a neighbourhood $U(x_0)$ of x_0, let $h \in V$ such that $[x,x_0+h] \subset U(x_0)$, then there exists $\theta \in]0,1[$ such that

$$f(x_0+h) = f(x_0)+Df(x_0)+\frac{1}{2}D^2f(x_0+\theta h).(h,h) \qquad \textit{(Taylor-Maclaurin formula)}\ .$$

If $f \in C^2(\Omega)$ and W is a Banach space then

$$f(x_0+h) = f(x_0) + Df(x_0).h + \int_0^1 [(1-t)Df(x_0+th).(h,h)]dt$$

(Taylor formula with integral remainder) .

Remark. *Taylor-Young's formula of order 1 is just the differentiability at x_0, while Taylor-Young's formula of order 2 is not equivalent to the twice differentiability. If $V=\mathbb{R}^n$ and $W=\mathbb{R}$, these formulas rewrite in terms of gradient and Hessian as*

$$f(x_0+h) = f(x_0) + \langle \nabla f(x_0), h\rangle + o(\|h\|),$$

and

$$f(x_0+h) = f(x_0) + \langle \nabla f(x_0), h\rangle + \frac{1}{2}\langle \mathrm{Hess}\, f(x_0)h, h\rangle + o(\|h\|^2),$$

where $\mathrm{Hess}\, f(x_0)$ is the Hessian if f at x_0, that is the symmetric matrix defined as

$$\mathrm{Hess}\, f(x_0) = (\partial_{ij} f(x_0))_{ij} \text{ with } \partial_{ij} f(x_0) = \partial_i(\partial_j f)(x_0) = D^2f(x_0)(e_i,e_j).$$

4. Optimality conditions

In this section we use differential calculus to give optimality conditions satisfied by the minimizers. We first consider the unconstrained case, i.e., we consider the case where the minimization set A is **open**.

Proposition 4.1. *(Necessary conditions) Let $J : A \subset V \to \mathbb{R}$ and $x_* \in A$. Assume that A is open and that J is differentiable at x_*. If x_* is a local minimizer of J in A then,*

- *First order condition:*

$$DJ(x_*) = 0, \ \textit{(Euler's equation)}. \tag{4.1}$$

- *Second order condition: If J is twice differentiable at x_* then, for all $h \in V$,*

$$D^2J(x_*).(h,h) \geq 0. \tag{4.2}$$

Remark. *If V is finite dimensional, (4.2) exactly means that the Hessian* Hess $J(x_*)$ *is positive semi-definite.*

Proof. This result relies on the classical 1-dimensional results applied in any admissible direction around x_*. As we assume that A is open, every direction is admissible. Indeed, let $h \in V$ and define

$$\begin{array}{rccc} \phi : & [-\epsilon, \epsilon] & \to & \mathbb{R} \\ & t & \mapsto & J(x_* + th). \end{array}$$

The application ϕ is differentiable at 0 and

$$\phi'(0) = DJ(x_*).h.$$

Moreover, 0 is a local minimizer of ϕ, thus $\phi'(0) = 0 = DJ(x_*).h$. Consequently, as every direction h is admissible, $DJ(x_*) = 0$. If now J is twice differentiable at x_*, then ϕ is twice differentiable at 0 and

$$\phi''(0) = D^2J(x_*).(h,h).$$

As x_* is a relative minimum of J, then 0 is a relative minimum of ϕ and $\phi''(0) \geq 0$. □

When the set A is not open, we have to restrict ourselves to admissible directions in which it is possible to make small variations form x_* .

Proposition 4.2. *Let U be an open neighbourhood of A in V, $J : U \to \mathbb{R}$ and $x_* \in A$. Assume that J is differentiable at x_*. If x_* is a local minimizer of J in A then*

$$DJ(x_*).h \geq 0$$

for any $h \in V$ such that $[x_*, x_* + h] \subset A$.

Proof. As J is differentiable at x_*, for $h \in V$ such that $[x_*, x_* + h] \subset A$ and for $t > 0$ small enough, $J(x_* + th) \geq J(x_*)$ and thus

$$DJ(x_*).h = \lim_{t \to 0} \frac{J(x_* + th) - J(x_*)}{t} \geq 0$$

□

When the set A is convex, Proposition 4.2 is known as Euler inequality and rewrites

Proposition 4.3. *(Euler inequality) Let U be an open neighbourhood of A in V, $J : U \to \mathbb{R}$ and $x_* \in A$. Assume that A is convex and that J is differentiable at x_*. If x_* is a local minimizer of J in A then for all $x \in A$,*

$$DJ(x_*)(x - x_*) \geq 0.$$

Proof. As A is convex, for all $x \in A$, $[x_*, x] \subset A$ and it is possible to apply Proposition 4.2 to the direction $x - x_* \in V$. □

Compare with the optimality condition found in Proposition 2.10.

4.1. **Sufficient conditions.** We now look for conditions insuring that a candidate x_* is a relative minimum of J. We thus assume that x_* satisfies the necessary first order condition $DJ(x_*) = 0$ (Euler equation). Such a point is called a *critical point*. It is well-known that this is not a sufficient condition to be a relative minimum (0 is a critical point of $t \mapsto t^3$). It is also well-known that the second order necessary condition (4.2) is not sufficient as well (0 satisfies this condition for $x \mapsto x^5$).

Theorem 4.4. *Let $J : A \subset V \to \mathbb{R}$ and $x_* \in A$. Assume that A is open and that x_* is a critical point of J, that is J is differentiable at x_* and $DJ(x_*) = 0$.*

- *If J is twice differentiable at x_* and if there exists $\alpha > 0$ such that for all $h \in V$,*

$$D^2J(x_*).(h, h) \geq \alpha \|h\|^2, \tag{4.3}$$

  then x_* is a strict relative minimum for J.
- *If J is twice differentiable in a neighbouring ball B centred at x_* and satisfies,*

$$D^2J(x).(h, h) \geq 0, \quad \forall x \in B \text{ and } \forall h \in V, \tag{4.4}$$

  then x_* is a relative minimum for J.

Remark: Notice that if V is finite dimensional, condition (4.3) is equivalent to say that the Hessian $\nabla J^2(x_*)$ is positive definite.

Proof. The proof follows from Taylor-Young and Taylor-Maclaurin formulas. □

Exercise 8. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = x^4 + y^4 - 4xy$. Study the critical points of f.*

Unfortunately, as well as the necessary conditions stated in the previous section are not sufficient, these sufficient conditions are not necessary. Indeed, the function $f(x) = x^4$ has a minimum at 0 but $f''(0) = 0$ so the first sufficient condition (4.3) is not true. To see that the second sufficient condition (4.4) is not necessary we can examine the function

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2 \left(2 + \sin \frac{1}{x}\right) \text{ if } x \neq 0 \text{ and } f(0) = 0.$$

Nevertheless, there is a particular but very important case where being a critical point is both necessary and sufficient to be a relative extremum: this is the case when the function J is convex.

4.2. **The case of convex functions.**

Proposition 4.5. *Let $A \subset V$ be an open set and $J : A \to \mathbb{R}$ be differentiable. Then*

- *J is convex in A iff for all $x, y \in A$, $J(y) \geq J(x) + DJ(x)(y-x)$,*
- *J is strictly convex in A iff for all $x, y \in A$, $x \neq y$, $J(y) > J(x) + DJ(x)(y-x)$.*

If moreover J is twice differentiable, then

- *J is convex iff for all $x \in A, h \in V$, $D^2J(x).(h,h) \geq 0$,*
- ***If** for all $x \in A, h \in V \setminus \{0\}$, $D^2J(x).(h,h) > 0$ then J is strictly convex.*

Order 1 characterization means that the graph of J is (strictly) above its tangent hyperplane everywhere. To see that the second order sufficient condition for strict convexity is not necessary we can take $f(x) = x^4$ defined in $\mathbb{R}$, it is strictly convex in $\mathbb{R}$ but $f''(0) = 0$.

Theorem 4.6. *Let V be a vector space and $A \subset V$ be a convex set. Let $J : A \to \mathbb{R}$ be a convex function. Then,*

- *if J is differentiable in a neighbourhood of A, then the minimizers of J are exactly the critical points of J. It means that J has a minimum in x_* if and only if for every $x \in A$*
$$DJ(x_*)(x - x_*) \geq 0;$$
- *if J has a relative minimum then it is a global minimum of J;*
- *if moreover J is strictly convex, then it has at most one minimizer and it is a strict minimizer.*

The proof of uniqueness when J is strictly convex is the following: assume, by contradiction, that x_1, x_2 are two distinct minimizers. Then, as J is strictly convex, we have $J\left(\frac{x_1+x_2}{2}\right) < \frac{1}{2}\left(J(x_1) + J(x_2)\right)$ which is a contradiction.

5. Minimization with constraints

In this section we focus on the case where A is not open and is defined by some constraints. We will consider

- equality constraints
- inequality constraints.

5.1. **Equality constraints.** In this section, we assume that

$$A = \{x \in V; g_1(x) = \cdots = g_p(x)\}.$$

for some functions $g_1, \cdots, g_p : V \to \mathbb{R}$. The characterization of the solution of this minimization problem with equality constraints (when it exists) relies on the implicit functions theorem and simple linear algebra.

Theorem 5.1. *(constrained extrema) Let V be a Banach space and $J, g_1, \cdots, g_p : V \to \mathbb{R}$. Let $x_* \in A$, assume that J, g_i are C^1 in a neighbourhood of x_* and that x_* is a relative minimum of J in A. Assume moreover that*

$$\text{the vectors } Dg_1(x_*), \cdots, Dg_p(x_*) \text{ are linearly independent.} \tag{5.1}$$

Then, there exist Lagrange multipliers $\lambda_1, \cdots, \lambda_p \in \mathbb{R}$ such that

$$DJ(x_*) + \sum_{i=1}^{p} \lambda_i Dg_i(x_*) = 0. \tag{5.2}$$

We begin with a glance at the linear case[5] where J is a continuous linear form on V and where the constraints are actually an intersection of hyperplanes. In this case, g_i are continuous linear forms so that J, g_i are C^1 in V and for all $x \in V$, $Dg_i(x) = g_i$ and $DJ(x) = J$. Therefore, in this case (5.2) rephrases as

$$J + \sum_{i=1}^{p} \lambda_i g_i = 0. \tag{5.3}$$

Moreover, in this particular case A is still a vector space, so that we can apply the first order condition of local minimality to any direction $h \in A$, that is,

$$\forall h \in \bigcap_{i=1}^{p} \ker g_i, \quad DJ(x).h = J(h) = 0 \Leftrightarrow \bigcap_{i=1}^{p} \ker g_i \subset \ker J. \tag{5.4}$$

Hence the linear case amounts to prove that (5.4) implies (5.3). This is a simple algebraic result which is a key ingredient in the proof of Theorem 5.1.

Proposition 5.2. *Let J, $g_1, \cdots, g_p : V \to \mathbb{R}$ be linear forms. Assume that*

$$\bigcap_{i=1}^{p} \ker g_i \subset \ker J,$$

then J is a linear combination of $g_1, \cdots, g_p$.

Proof. Let us define the linear application $F = (J, g_1, \cdots, g_p) : V \to \mathbb{R}^{p+1}$. By assumption $a := (1, 0, \cdots, 0) \notin \mathrm{Im}F$ so that F is not surjective and $\mathrm{Im}F$ is a subspace of $\mathbb{R}^{p+1}$ of codimension at least 1. There exists H hyperplane of $\mathbb{R}^{p+1}$ containing $\mathrm{Im}F$ and not a. Then, there exists $\Lambda = (\lambda_0, \lambda_1, \cdots, \lambda_p) \in \mathbb{R}^{p+1}$ such that for all $h \in V$, $\langle \Lambda, F(h) \rangle = 0$. Hence, for all $h \in V$,

$$\lambda_0 J(h) + \sum_{i=1}^{p} \lambda_i g_i(h) = 0 \Rightarrow \lambda_0 J + \sum_{i=1}^{p} \lambda_i g_i = 0.$$

As $a \notin H$, $\lambda_0 = \langle \Lambda, a \rangle \neq 0$. Thus we can divide by λ_0 and obtain the result. □

We are now going to give two ways of ending the proof. One when V has finite dimension, which allows to see A as a sub-manifold. On one hand, it is then easy to prove that $DJ(x_*)$ restricted to the tangent plane to A at x_* must be zero, and on the other hand, the tangent plane expresses directly from the differential of the constraints. Those two facts are enough to lead to a relation of the form (5.4) and conclude with Proposition

[5] This is an important idea in analysis to use differential calculus to to reduce oneself to a linear case.

5.2. We then give a less geometric proof in the general case where V is a Banach space, relying on the inverse function theorem.

Proof of Theorem 5.1 in finite dimension. Let $h \in T_{x_*}Dg := \{h \ ; \ Dg(x_*).h = 0\}$, then, thanks to assumptions of Theorem 5.1, there exists a differentiable arc $\gamma :]-\epsilon, \epsilon[\to M$ such that $\gamma(0) = x_*$ and $\gamma'(0) = h$. Thus, the one variable function $J \circ \gamma$ has a relative minimum at 0. Hence $(J \circ \gamma)'(0) = 0$ i.e., $DJ(\gamma(0)).(\gamma'(0)) = \langle \nabla J(x_*), h\rangle = 0$.

It is then clear that

$$T_{x_*}Dg = \ker Dg(x_*) = \bigcap_{i=1}^{p} \ker Dg_i(x_*),$$

and applying Proposition **??** leads to $\cap_{i=1}^{p} \ker Dg_i(x_*) \subset \ker DJ(x_*)$ and we conclude with Proposition 5.2 that $DJ(x_*) \in \operatorname{span}(Dg_1(x_*), \cdots, Dg_p(x_*)$. □

Before proving Theorem 5.1 in a Banach space we start by recalling the inverse function theorem in Banach spaces.

Theorem 5.3. *(Inverse function theorem) Let V, W be Banach spaces, let $\Omega \subset V$ be an open set and $f : \Omega \to W$ be a function of class C^1. Assume that for some $x \in \Omega$, $Df(x)$ is an isomorphism, then there exists an open set $\mathcal{V} \subset \Omega$ containing x and an open set $\mathcal{W} \subset W$ containing $f(x)$ such that f is a C^1-diffeomorphism when restricted from $\mathcal{V}$ to $\mathcal{W}$.*

Proof of Theorem 5.1 in a Banach space. First of all, it is possible to assume that $x_* = 0$ without loss of generality[6]. Let $\mathcal{U} \subset V$ be an open ball centred at 0 and such that

$$\forall x \in \mathcal{U} \cap A, \quad J(x) \geq J(0).$$

Define $F : \mathcal{U} \to \mathbb{R}^{p+1}$ s.t. for $x \in \mathcal{U}, F(x) = (J(x), g_1(x), \cdots, g_p(x))$. Note that for all $c < J(0)$, $(c, 0, \cdots, 0) \notin F(\mathcal{U})$ but $F(0) = (J(0), 0, \cdots, 0) \in F(\mathcal{U})$, whence $F(\mathcal{U})$ cannot contain any open set around $F(0)$. Assume, towards a contradiction that $DF(0) \in \mathcal{L}(V, \mathbb{R}^{p+1})$ were surjective. Note that $G = \ker DF(0)$ is a (closed since F is C^1) vector subspace of V. Take E a supplementary of G in V so that E is isomorphic to $\mathrm{Im}DF(0) = \mathbb{R}^{p+1}$ and $DF(0)_{|E} \in \mathrm{Isom}(E, \mathbb{R}^{p+1})$ (isomorphism theorem). Define $\tilde{F} = F_{|E} : \mathcal{U} \cap E \to \mathbb{R}^{p+1}$, for all $x \in \mathcal{U} \cap E$, $\tilde{F}(x) = F(x)$ as the restriction of F to E and equip E with the norm $\|\cdot\|_E$ induced by V. Then $\tilde{F}$ is differentiable and $D\tilde{F}(x) = DF(x)_{|E} \in \mathcal{L}(E, \mathbb{R}^{p+1})$. Indeed, let $x \in \mathcal{U} \cap E$ and $h \in E$,

$$\begin{aligned} \tilde{F}(x+h) - \tilde{F}(x) &= F(x+h) - F(x) \\ &= DF(x).h + o(\underbrace{\|h\|_V}_{=\|h\|_E}). \end{aligned}$$

Notice that, as E and $\mathbb{R}^{p+1}$ are finite dimensional, E could actually be equipped with any norm. The function $\tilde{F} : \mathcal{U} \cap E \to \mathbb{R}^{p+1}$ is differentiable and of class C^1 since DF is

[6] Define $\hat{\Omega} = \Omega - x_*$, $\hat{J} : \hat{\Omega} \to \mathbb{R}, \hat{J}(x) = J(x + x_*)$ and $\hat{g}_i : \hat{\Omega} \to \mathbb{R}$, $\hat{g}_i(x) = g_i(x + x_*)$. Then $D\hat{J}(0) = DJ(x_*)$, $D\hat{g}_i(0) = Dg_i(x_*)$ and minimizing J in Ω is equivalent to minimizing $\hat{J}$ in $\hat{\Omega}$ with constraint $\hat{g}_i = 0$.

continuous in $\mathcal{U}$ and for $x, y \in \mathcal{U} \cap E$,

$$\begin{aligned}\|D\tilde{F}(x) - D\tilde{F}(y)\|_{\mathcal{L}(E,\mathbb{R}^{p+1})} &= \sup_{h\in E, \|h\|_E=1} \|D\tilde{F}(x).h - D\tilde{F}(y).h \\ &= \sup_{h\in E, \|h\|_V=1} \|DF(x).h - DF(y).h\| \\ &\leq \sup_{h\in V, \|h\|_V=1} \|DF(x).h - DF(y).h\| \\ &= \|DF(x) - DF(y)\|_{\mathcal{L}(V,\mathbb{R}^{p+1})}.\end{aligned}$$

Moreover, $0 \in \mathcal{U} \cap E$ and $D\tilde{F}(0) : E \to \mathbb{R}^{p+1}$ is an isomorphism, by the inverse function theorem, there exists $\mathcal{U}' \subset E \cap \mathcal{U}$ open in E containing 0 and $\mathcal{W} \subset \mathbb{R}^{p+1}$ open set containing $\tilde{F}(0) = F(0)$ such that $\tilde{F} : \mathcal{U}' \to \mathcal{W}$ is a C^1 diffeomorphism. Consequently,

$$F(0) \in \mathcal{W} = \tilde{F}(\mathcal{U}') \subset \tilde{F}(\mathcal{U}) = F(\mathcal{U} \cap E) \subset F(\mathcal{U}),$$

and this contradicts the fact that there is no open subset of $F(\mathcal{U})$ containing $F(0)$ and thus $DF(0)$ is not surjective. Eventually, $\mathrm{Im}DF(0)$ is a strict subset of $\mathbb{R}^{p+1}$ and is therfore contained in some hyperplane of $\mathbb{R}^{p+1}$. There exists $\Lambda = (\lambda_0, \lambda_1, \cdots, \lambda_p) \in \mathbb{R}^{p+1}$ such that for all $h \in V$, $\langle \Lambda, DF(0)\rangle = 0$, i.e.,

$$\lambda_0 DJ(0) + \sum_{i=1}^{p} \lambda_i Dg_i(0) = 0.$$

As $\lambda_0 = 0$ would imply that $(Dg_1(0), \cdots, Dg_p(0))$ are linearly dependent, we conclude that $\lambda_0 \neq 0$. We can divide by λ_0 to obtain the conclusion. □

Exercise 9. *Maximize in $\mathbb{R}^2$ the cost function $J(x, y) = x^4 + y^4$ under the constraint $x^6 + y^6 = 1$.*

Exercise 10. *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x, y, z) = x - y + z$. Find the extrema of f under the constraints $x^2 + y^2 + z^2 = 4$ and $x + y + z = 1$.*

Exercise 11. *(An example in infinite dimension) Let us consider the vector space*

$$E_0 = \{u \in C^2([a, b], \mathbb{R}); u(a) = u(b) = 0\},$$

provided with the C^2-norm $\|u\|_{C^2} = \sup_{[a,b]}(|u| + |u'| + |u''|)$. Let $L, K : \mathbb{R}^3 \to \mathbb{R}$ be C^2 and for all $u \in E_0$, we define

$$J(u) = \int_a^b L(t, u(t), u'(t))dt \text{ and } G(u) = \int_a^b K(t, u(t), u'(t))dt.$$

i) *Show that J is differentiable and compute its differential.*

ii) *Show that if u_* minimizes J in E_0 then for all $t \in [a, b]$,*

$$\frac{d}{dt}\left(\partial_3 L(t, u(t), u'(t))\right) = \partial_2 L(t, u(t), u'(t)).$$

iii) *Show that if u_* minimizes J in E_0 under the constraint $G(u) = \alpha$ then there exists $\lambda \in \mathbb{R}$ such that for all $t \in [a, b]$,*

$$\frac{d}{dt}\left(\partial_3 (L + \lambda K)(t, u(t), u'(t))\right) = \partial_2 (L + \lambda K)(t, u(t), u'(t)).$$

iv) *Show that if L, K are autonomous (independent of x_1) then u_* satisfies Erdmann's condition: there exists a constant $\mu \in \mathbb{R}$ such that for all $t \in [a, b]$,*

$$(L + \lambda K)(u(t), u'(t)) - u'(t)\partial_2(L + \lambda K)(u(t), u'(t)) = \mu.$$

v) *Apply this result to Dido's problem: given two points A and B, determine the curve (of fixed length) joining these two points and such that the area enclosed by the curve and the segment $[A, B]$ is maximal. In particular show that such a curve has constant curvature.*

5.2. **Inequality constraints.** In this section V is a Hilbert space. We assume that the minimization set is given by

$$A = \{x \in V; h_1(x) \leq 0, \cdots, h_p(x) \leq 0\},$$

for some C^1 functions $h_1, \cdots, h_p : V \to \mathbb{R}$.

Definition 5.4. Let $x \in A$, the set $I(x) = \{i \in \{1, \cdots, q\}; h_i(x) = 0\}$ is called the set of *active constraints* at x.

Definition 5.5. The constraints are said to be *qualified* at $x \in A$ if there exists a direction $\tilde{w} \in V$ such that for all $i \in I(x)$

$$\text{either } \langle \nabla h_i(x), \tilde{w} \rangle < 0$$
$$\text{or } \langle \nabla h_i(x), \tilde{w} \rangle = 0 \text{ and } h_i \text{ is affine.}$$

Theorem 5.6. *(Karush-Kuhn-Tucker) Let V be a Hilbert space, $J, h_1, \cdots, h_q : V \to \mathbb{R}$. We assume that $x_* \in A$, J, h_j are C^1 in a neighbourhood of x_* and that x_* is a local minimum of J in A. If the constraints are qualified at x_*, then there exist Lagrange multipliers $\mu_1, \cdots, \mu_q \in \mathbb{R}^+$ such that*

$$DJ(x_*) + \sum_{j=1}^{q} \mu_j Dh_j(x_*) = 0$$

and $\mu_j = 0$ if $h_j(x_*) = 0$.

Proof. We consider the set

$$\tilde{A}(x_*) = \{w \in V; \langle \nabla h_j(x_*), w \rangle \leq 0, \ \forall i \in I(x_*)\},$$

(this is the set of "admissible directions"). Let $\tilde{w}$ be such that $\langle \nabla h_j(x_*), \tilde{w} \rangle < 0$ if h_j is not affine and $j \in I(x_*)$, or $\langle \nabla h_j(x_*), \tilde{w} \rangle = 0$ if h_j is affine and $j \in I(x_*)$. Let $\delta > 0$. We will show that $x_* + \epsilon(w + \delta\tilde{w}) \in A$ for ϵ sufficiently small. We distinguish three cases:

1) If $j \notin I(x_*)$, then $h_j(x_*) < 0$ and $h_j(x_* + \epsilon(w + \delta\tilde{w})) < 0$ by continuity of h_j if ϵ is sufficiently small.
2) If $j \in I(x_*)$ and $\langle \nabla h_j(x_*), \tilde{w} \rangle < 0$, then

$$\begin{aligned} h_j(x_* + \epsilon(w + \delta\tilde{w})) &= h_j(x_*) + \epsilon\langle \nabla h_j(x_*), w + \delta\tilde{w} \rangle + o(\epsilon) \\ &= 0 + \epsilon \underbrace{\langle \nabla h_j(x_*), w \rangle}_{\leq 0} + \epsilon\delta \underbrace{\langle \nabla h_j(x_*), \tilde{w} \rangle}_{<0} + o(\epsilon) \\ &\leq 0 \text{ for } \epsilon \text{ sufficiently small .} \end{aligned}$$

3) If $j \in I(x_*)$ and $\langle \nabla h_j(x_*), \tilde{w} \rangle = 0$ then h_j is affine and

$$\begin{aligned} h_j(x_* + \epsilon(w + \delta\tilde{w})) &= h_j(x_*) + \epsilon\langle \nabla h_j(x_*), w + \delta\tilde{w} \rangle \\ &= \epsilon\langle \nabla h_j(x_*), w \rangle \leq 0. \end{aligned}$$

Now if x_* is a local minimizer of J in A we have

$$\frac{J(x_* + \epsilon(w + \delta\tilde{w})) - J(x_*)}{\epsilon} \geq 0$$

and passing to the limit as ϵ goes to 0 we find

$$\langle DJ(x_*), w + \delta\tilde{w}\rangle \geq 0 \quad \forall w \in \tilde{A}(x_*),\ \forall \delta > 0.$$

We let δ go to zero and obtain $\langle DJ(x_*), w\rangle \geq 0$ for all $w \in A(x_*)$. This means that if w is such that $\langle \nabla h_j(x_*), w\rangle \leq 0$ for all $i \in I(x_*)$ then $\langle DJ(x_*), w\rangle \geq 0$. We conclude by an algebraic lemma known as the Farkas lemma stated below. □

Lemma 5.7. *Let $M \in \mathbb{N}^*$ a_j, $1 \leq j \leq M$, be elements of V Hilbert space and $b \in V$. Then*

$$\{w \in V; \langle a_j, w\rangle \geq 0,\ 1 \leq j \leq M\} \subset \{w \in V; \langle b, w\rangle \geq 0\}$$

holds if and only if there exist $\lambda_j \geq 0$, $1 \leq j \leq M$ such that

$$b = \sum_{j=1}^{M} \lambda_j a_j.$$

For the proof we refer to [5, Theorem 9.1-1].

Definition 5.8. If the functions h_j, $1 \leq j \leq M$ are **convex**, then we say that the constraints are *qualified* at x_* if there exists $\tilde{v} \in V$ such that for all $j \in \{1, \cdots, q\}$

* either $h_j(\tilde{v}) < 0$

* or $h_j(\tilde{v}) = 0$ and h_j is affine.

Proposition 5.9. *If the constraints are qualified in the sense of Definition 5.8 they are also qualified in the sense of Definition 5.5.*

Proof. If $j \in I(x_*)$ and $h_j(\tilde{v}) < 0$ then, since h_j is convex

$$\langle \nabla h_j(x_*), \tilde{v} - x_*\rangle = h_j(x_*) + \langle \nabla h_j(x_*), \tilde{v} - x_*\rangle \leq h(\tilde{v}) < 0.$$

If $i \in I(x_*)$ and $h_j(\tilde{v}) = 0$ then h_j is affine and

$$\langle \nabla h_j(x_*), \tilde{v} - x_*\rangle = h_j(\tilde{v}) - h_j(x_*) = 0.$$

In both cases we can take $\tilde{w} := \tilde{v} - x_*$ and this $\tilde{w}$ satisfies the properties of the first definition. □

Theorem 5.10. *(KKT-convex case) Let V be a Hilbert space, $J, h_1, \cdots, h_q : V \to \mathbb{R}$. We assume that $x_* \in A$, J, h_j are C^1 in a neighbourhood of x_* and that x_* is a local minimum of J in A. We assume furthermore that $J, h_j : V \to \mathbb{R}$ are* ***convex****. If the constraints are qualified at x_* (in the sense of Definition 5.8), then there exist Lagrange multipliers $\mu_1, \cdots, \mu_q \in \mathbb{R}^+$ such that*

$$DJ(x_*) + \sum_{j=1}^{q} \mu_j Dh_j(x_*) = 0$$

and $\sum_{j=1}^{q} \mu_j h_j(x_*) = 0$.
Conversely, if $J, h_j : V \to \mathbb{R}$ are ***convex*** *and if there exist $\mu_i, 1 \leq i \leq q$ such that the KKT conditions are verified at x_* then x_* is a minimum of J in A.*

5.3. **Equality and inequality constraints.** Here we assume that V is a finite dimensional vector space and that

$$A = \{x \in V; g_1(x) = \cdots = g_p(x) = 0,\ h_1(x) \leq 0, \cdots, h_q(x) \leq 0\},$$

for some C^1 functions $g_1, \cdots, g_p, h_1, \cdots, h_q : V \to \mathbb{R}$.

Definition 5.11. We say that the constraints are *qualified* at x_* if $(\nabla g_1(x_*), \cdots, \nabla g_p(x_*))$ are linearly independent and if there exists a direction $\tilde{w} \in \bigcap_{i=1}^{p} (\nabla g_i(x_*))^{\perp}$ such that for all $j \in I(x_*)$, $\langle \nabla h_j(x_*), \tilde{w} \rangle < 0$.

Remark: A stronger assumption implying that the constraint are qualified at x_* is: $(\nabla g_1(x_*), \cdots, \nabla g_p(x_*), \nabla h_1(x_*), \cdots, \nabla h_q(x_*))$ are linearly independent.

Theorem 5.12. *(KKT) Let $J, g_1, \cdots, g_p, h_1, \cdots, h_q : V \to \mathbb{R}$. Let $x_* \in A$, assume that J, g_i, h_j are C^1 in a neighbourhood of x_* and that x_* is a local minimum of J in A. If the constraints are qualified at $x_* \in A$ then there exist Lagrange multipliers $\lambda_1, \cdots, \lambda_p, \mu_1, \cdots, \mu_q \in \mathbb{R}$ such that*

$$\nabla J(x_*) + \sum_{i=1}^{p} \lambda_i \nabla g_i(x_*) + \sum_{j=1}^{q} \mu_j \nabla h_j(x_*) = 0, \tag{5.5}$$

$$\mu_j \geq 0 \quad \textit{for all } j = 1, \cdots, q, \tag{5.6}$$

$$\mu_j = 0 \quad \textit{if } h_j(x_*) = 0. \tag{5.7}$$

Exercise 12. *Minimize $f(x, y) = -x + y$ under the constraints $y \geq x^2$ and $x + y \leq 1$.*

Correction: We define $h_1(x, y) = x^2 - y$, $h_2(x, y) = x + y - 1$ and

$$A := \{(x, y) \in \mathbb{R}^2; h_1(x, y) \leq 0,\ h_2(x, y) \leq 0\}.$$

Existence: The infimum is achieved because A is compact (it is closed and bounded) and f is continuous.

Qualification of the constraints: We have $\nabla h_1(x, y) = (2x, -1) \neq 0$ and $\nabla h_2(x, y) = (1, 1) \neq 0$. Therefore, when only one constraint i_0 is active, the qualification is satisfied since the family $(\nabla h_{i_0}(x, y))$ is linearly independent. If both constraints are active, then

$h_1(x,y) = h_2(x,y) = 0$. The solutions of these equations are

$$\left(\frac{-1-\sqrt{5}}{2}, \frac{3+\sqrt{5}}{2}\right), \left(\frac{-1+\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2}\right)$$

and at those two points ∇h_1 and ∇h_2 are independent. Thus the constraints are qualified at every point of A.

KKT conditions: If (x,y) minimizes f in A, there exist $\lambda \geq 0, \mu \geq 0$ such that

$$\left\{ \begin{array}{r} \nabla f(x,y) + \lambda \nabla h_1(x,y) + \mu \nabla h_2(x,y) = 0 \\ \lambda h_1(x,y) = 0 \text{ and } \mu h_2(x,y) = 0 \\ h_1(x,y) \leq 0 \text{ and } h_2(x,y) \leq 0 \end{array} \right. \iff \left\{ \begin{array}{r} -1 + 2\lambda x + \mu = 0 \\ 1 - \lambda + \mu = 0 \\ \lambda(x^2 - y) = 0 \\ \mu(x+y-1) = 0 \\ x^2 \leq y \text{ and } x + y \leq 1 \end{array} \right. .$$

* If $\lambda = 0$, then $1 = \mu = -1$ which is impossible.
* If $\lambda \neq 0$ and $\mu = 0$, then $\lambda = 1$ and then $x = 1/2$ and $y = x^2 = 1/4$. We do have $(1/2, 1/4) \in A$ and $f\left(\frac{1}{2}, \frac{1}{4}\right) = -\frac{1}{4}$.
* If $\lambda \neq 0$ and $\mu \neq 0$, then $(x,y) \in \left\{ \left(\frac{-1-\sqrt{5}}{2}, \frac{3+\sqrt{5}}{2}\right), \left(\frac{-1+\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2}\right) \right\}$ and

$$f\left(\frac{-1-\sqrt{5}}{2}, \frac{3+\sqrt{5}}{2}\right) = 2 + \sqrt{5} > -\frac{1}{4} \text{ and } f\left(\frac{-1+\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2}\right) = 2 - \sqrt{5}.$$

As $2 - \sqrt{5} + \frac{1}{4} = \frac{9-4\sqrt{5}}{4} > 0$ (since $9 - 4\sqrt{5} = 9 - \sqrt{80} > 0$, we have $2 - \sqrt{5} > -1/4$. We conclude that there exists a unique minimizer of f in A and it is $(\frac{1}{2}, \frac{1}{4})$.

6. Numerical algorithms: descent methods

6.1. Elliptic functionals.

Definition 6.1. (Elliptic functional) A functional $J : V \to \mathbb{R}$ defined on a Hilbert space $(V, \langle \cdot, \cdot \rangle)$ is called *elliptic* if it is C^1 and if there exists $\alpha > 0$ such that

$$\langle \nabla J(y) - \nabla J(x), y - x \rangle \geq \alpha \|y - x\|^2 \quad \forall x, y \in V.$$

Proposition 6.2. *A C^2 functional $J : V \to \mathbb{R}$ is elliptic iff*

$$\langle \text{Hess}\, J(x) y, y \rangle \geq \alpha \|y\|^2 \quad \forall x, y \in V.$$

Proof. If J is C^2 and elliptic we have

$$\begin{aligned} \langle \text{Hess}\, J(x) y, y \rangle &= \lim_{t \to 0} \frac{\langle \nabla J(x + ty) - \nabla J(x), y \rangle}{t} \\ &= \lim_{t \to 0} \frac{\langle \nabla J(x + ty) - \nabla J(x), ty \rangle}{t^2} \geq \alpha \|y\|^2. \end{aligned}$$

For the converse we apply Taylor-Maclaurin formula to

$$f : w \in V \mapsto \langle \nabla J(w), y - x \rangle$$

with fixed $x, y \in V$. We have

$$\begin{aligned}\langle \nabla J(y) - \nabla J(x), y - x\rangle &= f(y) - f(x)\\ &= Df(x + \theta(y-x)).(y-x) \text{ for some } 0 < \theta < 1,\\ &= \langle \text{Hess}\, J(x + \theta(y-x)), y - x\rangle \geq \alpha \|x-y\|^2.\end{aligned}$$

□

Lemma 6.3. *Let J be α-elliptic, then J is coercive, strictly convex and satisfies*

$$J(y) - J(x) \geq \langle \nabla J(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \quad \forall x, y \in V.$$

Proof. We apply Taylor formula with integral remainder:

$$\begin{aligned}J(y) - J(x) &= \int_0^1 \langle \nabla J(x + t(y-x), y - x\rangle dt\\ &= \langle \nabla J(x), y - x\rangle + \int_0^1 \left[\langle \nabla J(x + t(y-x)), y-x\rangle - \langle \nabla J(x), y - x\rangle\right] dt\\ &= \langle \nabla J(x), y - x\rangle + \int_0^1 \left[\langle \nabla J(x + t(y-x)) - \nabla J(x), y - x\rangle\right] \frac{dt}{t}\\ &\geq \langle \nabla J(x), y - x\rangle + \int_0^1 \alpha t \|y - x\|^2 dt\\ &\geq \langle \nabla J(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2.\end{aligned}$$

This means that J is strictly convex thanks to the characterization of first order of convex functions (cf. Proposition 4.5). We also see that J is coercive since

$$\begin{aligned}J(y) &\geq J(0) + \langle \nabla J(0), y\rangle + \frac{\alpha}{2}\|y\|^2\\ &\geq J(0) - \|\nabla J(0)\|\|y\| + \frac{\alpha}{2}\|y\|^2\\ &\to +\infty \text{ when } \|y\| \to +\infty.\end{aligned}$$

□

Theorem 6.4. *Let $J : V \to \mathbb{R}$ be a C^1 elliptic functional. Let $A \subset V$ be a closed convex set. Then there exists a unique $x_* \in A$ such that*

$$J(x_*) = \min_{x \in A} J(x) > -\infty. \tag{6.1}$$

Furthermore x_* satisfies (6.1) iff $\langle \nabla J(x_*), x - x_*\rangle \geq 0$ for all $x \in A$.

Proof. Since J is coercive and strictly convex (and continuous) we can apply the existence result Theorem 2.17. The minimizer is unique by strict convexity, the optimality condition follows from Proposition 4.3. □

6.2. **General principle of descent methods.** In the following V is a finite dimensional vector space and we take $V = \mathbb{R}^n$. We consider the problem of finding $x_* \in V$ such that $J(x_*) = \inf_{x \in V} J(x)$. We want to construct a sequence $(x_k)_{k \in \mathbb{N}}$ such that $x_k \to x_*$ when $k \to +\infty$. The general principle of descent methods used to construct such sequences is the following

```
• Fix $x_0 \in \mathbb{R}^n$ ($k = 0$)
• while stopping criterion
    - choose $d_k \in \mathbb{R}^n$ a descent direction
    - choose $\rho_k$ a step size
    - set $x_{k+1} = x_k + \rho_k d_k$
    - do $k \leftrightarrow k + 1$.
```

Different choices of descent directions and of step sizes give rise to different descent methods that we will now examine.

As a first example we take a look at the *relaxation method.* Here the successive directions are taken to be the vectors of the canonical basis of $\mathbb{R}^n$: $(e_1, \cdots, e_n)$.

Algorithm 1: Relaxation method

```
• Choose $x_0 \in \mathbb{R}^n$
• Given $x_k$, construct $x_{k+1}$ as follows:
    - $x_{k+1,1} = x_k + \rho_{k,1} e_1$ with $\rho_{k,1}$ such that
          $J(x_k + \rho_{k,1} e_1) = \min_{\rho \in \mathbb{R}} J(x_k + \rho e_1);$
    - $x_{k+1,2} = x_{k+1,1} + \rho_{k,2} e_2$ with $\rho_{k,2}$ such that
          $J(x_{k+1,1} + \rho_{k,2} e_2) = \min_{\rho \in \mathbb{R}} J(x_{k+1,1} + \rho e_2);$
    - $\cdots$
    - $x_{k+1} = x_{k+1,n} = x_{k+1,n-1} + \rho_{k,n} e_n$ with $\rho_{k,n}$ such that
          $J(x_{k+1,n-1} + \rho_{k,n} e_n) = \min_{\rho \in \mathbb{R}} J(x_{k+1,n-1} + \rho e_n).$
```

Theorem 6.5. *If the functional $J : \mathbb{R}^n \to \mathbb{R}$ is elliptic then the relaxation method converges.*

For the proof we refer to Theorem 8.4-2 in [5].

6.3. Gradient methods.

6.3.1. *Principle.*

Definition 6.6. (Descent direction) We say that $d \in \mathbb{R}^n$ is a (strict) *descend direction* at x_0 if there exists $\rho_0 > 0$ such that for all $\rho \in]0, \rho_0[$

$$J(x_0 + \rho d) < J(x_0).$$

A descent direction is a direction along which J is locally decreasing: "if we move a little bit in this direction then J decreases."

Proposition 6.7. *Let $x, d \in \mathbb{R}^n$*

i) *if $\langle \nabla J(x), d \rangle < 0$ then d is a descent direction at x.*
ii) *If $-\nabla J(x) \neq 0$ then $-\nabla J(x)$ is a descent direction at x and it is even the steepest direction.*

Proof. i) Let $h : t \in \mathbb{R} \mapsto J(x + td)$. We have that $h'(t) = \langle \nabla J(x + td), d \rangle$ so that $h'(0) = \langle \nabla J(x), d \rangle < 0$. Moreover $h'(0) = \lim_{t \to 0} \frac{J(x+td) - J(x)}{t} < 0$, thus there exists $t_0 > 0$ such that for all $0 < t < t_0$, $J(x+td) - J(x) < 0$. This means that d is a descent direction.

ii) We remark that $\langle \nabla J(x), -\nabla J(x)\rangle = -|\nabla J(x)|^2 < 0$. Thus, by i) we deduce that $-\nabla J(x)$ is a descent direction. Besides $J(x+h) - J(x) = \langle \nabla J(x), h\rangle + o(\|h\|)$ and $|\langle \nabla J(x), h\rangle| \leq |\nabla J(x)||h|$ with equality iff $h = \lambda \nabla J(x)$. Hence $-\nabla J(x)$ is the steepest direction.

□

The gradient methods consist in choosing a descent direction $d_k = -\nabla J(x_k)$, the choice of the step size ρ_k is specific to each method.

6.3.2. *The gradient method with optimal step.* We first present the algorithm:

```
Algorithm 2: Gradient with optimal step
    • Take $x_0 \in \mathbb{R}$.
    • Take $\rho_k \in \mathbb{R}$ such that $J(x_k - \rho_k \nabla J(x_k)) = \inf_{\rho\in\mathbb{R}} J(x_k - \rho \nabla J(x_k))$.
    • Take $x_{k+1} = x_k - \rho_k \nabla J(x_k)$.
```

Theorem 6.8. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be an elliptic functional, then the gradient method with optimal step converges.*

Proof. i) Without loss of generality we can assume $\nabla J(x_k) \neq 0$ for every $k \geq 0$. Indeed if $\nabla J(x_{k_0}) = 0$, since J is strictly convex then x_{k_0} is the minimizer and the method converges. We define

$$\begin{array}{rlll} \varphi_k & : \mathbb{R} & \to & \mathbb{R} \\ & \rho & \mapsto & J(x_k - \rho \nabla J(x_k)), \end{array}$$

φ_k is also coercive and strictly convex, so it admits a unique minimizer ρ_k characterized by

$$\varphi'(\rho_k) = -\langle \nabla J(x_k - \rho_k \nabla J(x_k)), \nabla J(x_k)\rangle = 0.$$

It means that

$$\langle \nabla J(x_{k+1}), \nabla J(x_k)\rangle = 0 :$$

two successive directions are orthogonal. Since $x_{k+1} = x_k - \rho_k \nabla J(x_k)$ we also have

$$\langle \nabla J(x_{k+1}), x_{k+1} - x_k\rangle = 0.$$

Thus by using Deinition 6.1 we deduce that

$$J(x_k) - J(x_{k+1}) \geq \frac{\alpha}{2}\|x_k - x_{k+1}\|^2. \tag{6.2}$$

ii) We have that $(J(x_k))_k$ is decreasing (by construction) and bounded from below by $J(x_*)$, we deduce that

$$\lim_{k\to\infty} J(x_k) - J(x_{k+1}) = 0,$$

by (6.2) this implies that $\lim_{k\to\infty} \|x_k - x_{k+1}\| = 0$.

iii) By using orthogonality of two successive descent directions we can write

$$\begin{aligned} \|\nabla J(x_k)\|^2 &= \langle \nabla J(x_k), \nabla J(x_k) - \nabla J(x_{k+1})\rangle \\ &\underbrace{\leq}_{\text{Cauchy-Schwarz}} \|\nabla J(x_k)\|\|\nabla J(x_k) - \nabla J(x_{k+1})\|. \end{aligned}$$

Hence

$$\|\nabla J(x_k)\| \leq \|\nabla J(x_k) - \nabla J(x_{k+1})\|. \tag{6.3}$$

iv) Since $(J(x_k))_k$ is decreasing, the sequence $(x_k)_k$ is bounded because $J(x_k) \to +\infty$ as $\|x_k\| \to +\infty$ (coercivity). The differential DJ (or the gradient) is continuous, thus it is uniformly continuous on compact sets of $\mathbb{R}^n$, thus

$$\lim_{k\to+\infty} \|x_k - x_{k+1}\| = 0 \Rightarrow \lim_{k\to+\infty} \|\nabla J(x_k) - \nabla J(x_{k+1})\| = 0.$$

From (6.3) we deduce that

$$\lim_{k\to+\infty} \nabla J(x_k) = 0. \tag{6.4}$$

v) We write

$$\begin{aligned}\alpha\|x_k - x\|^2 &\leq \langle \nabla J(x_k) - \nabla J(x), x_k - x\rangle \\ &= \langle \nabla J(x_k), x_k - x\rangle \\ &\leq \|\nabla J(x_k)\|\|x_k - x\|.\end{aligned}$$

this implies that $\|x_k - x\| \leq \frac{1}{\alpha}\|\nabla J(x_k)\|$ and we can use (6.4) to conclude.

□

Recall:

- A function is continuous in U open set if

$$\forall x \in U,\ \forall \epsilon > 0,\ \exists \eta_x > 0 \ \text{ s.t. } \|x - y\| < \eta_x \Rightarrow \|f(x) - f(y)\| < \epsilon.$$

- A function f is uniformly continuous in U if

$$\forall \epsilon > 0,\ \exists \eta > 0 \text{ s.t. } \|x - y\| < \eta \Rightarrow \|f(x) - f(y)\| < \epsilon.$$

6.3.3. *The gradient method with fixed step size.* In this method we choose always the same ρ, hence we do not have to sole a $1D$ minimization problem at each step. However, for the convergence of the method we need to make a supplementary assumption on the cost function J.

Algorithm 3: Gradient method with fixed step

- Take $x_0 \in \mathbb{R}^n$ and choose$\rho > 0$
- Take $x_{k+1} = x_k - \rho\nabla J(x_k)$.

Theorem 6.9. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be an elliptic functional with elliptic constant $\alpha > 0$. Assume that there exists $M > 0$ such that for every $x, y \in \mathbb{R}^n$,*

$$\|\nabla J(y) - \nabla J(x)\| \leq M\|x - y\|. \tag{L}$$

Then, if $0 < \rho < \frac{2\alpha}{M^2}$ then

- *the gradient method with fixed step size ρ converges to the unique minimizer x_* of J,*
- *the convergence is of order 1:*

$$\|x_{k+1} - x_k\| \leq \beta\|x_k - x_*\|$$

with $\beta = \sqrt{1 - 2\alpha\rho + M^2\rho^2} < 1$.

Proof. As $\nabla J(x_*) = 0$ then

$$\begin{aligned} x_{k+1} - x_* &= x_k - \rho \nabla J(x_k) + \nabla J(x_*) - x_* \\ &= x_k - x_* + \rho(-\nabla J(x_k) + \nabla J(x_*)). \end{aligned}$$

Therefore,

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\rho\langle \nabla J(x_k) - \nabla J(x_*), x_k - x_*\rangle + \rho^2 \|\nabla J(x_k) - \nabla J(x_*)\| \\ &\leq (1 + \rho^2 M^2 - 2\rho\alpha)\|x_k - x_*\|^2. \end{aligned}$$

This implies that $\|x_k - x_*\| \leq \beta^k \|x_0 - x_*\|$ for $\beta = (1 + \rho^2 M^2 - 2\rho\alpha)^{1/2}$. □

Exercise 13. *Under the same assumptions as before, show that with variable step sizes ρ_k such that there exist $a, b > 0$ with $0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2}$ then the gradient method converges and the conclusions of the previous theorem hold with*

$$\beta = \max\{\sqrt{1 - 2\alpha a M^2 a^2}, \sqrt{1 - 2\alpha b + M^2 b^2}\}.$$

Remark: *(case of a quadratic functional) Let A be a $n \times n$ symmetric, positive definite matrix, $b \in \mathbb{R}^n$ and J be the quadratic functional defined in $\mathbb{R}^n$ as $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots \lambda_n$ be the eigenvalues of A then J satisfies the assumptions of Theorem 6.9 with $\alpha = \lambda_1$ and $M = \lambda_n$ and $\beta = \frac{2\lambda_1}{\lambda_n^2}$.*

6.3.4. *The conjugate gradient method (without constraint).* For $A \in S_n^{++}(\mathbb{R})$ and $b \in \mathbb{R}$, we consider $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ and the minimization problem

$$\inf_{x \in \mathbb{R}^n} J(x) = J(x_*).$$

Definition 6.10. We say that two non-zero directions $x, y \in \mathbb{R}^n$ are *A-conjugated* if $\langle x, Ay\rangle = \langle Ax, y\rangle = 0$.

Lemma 6.11. *We consider a family $d_0, \cdots, d_{n-1}$ of n A-conjugated vectors. Let $x_0 \in \mathbb{R}^n$, for $0 \leq k \leq n-1$ we define*

$$x_{k+1} = x_k + \lambda_k d_k,$$

where $J(x_k + \lambda_k d_k) = \inf_{\lambda \in \mathbb{R}} J(x_k + \lambda d_k)$. Then

$$\lambda_k = -\frac{\langle d_k, Ax_k - b\rangle}{\langle d_k, Ad_k\rangle}.$$

Proof. We compute $J(x_k + \lambda d_k) = J(x_k) + \lambda\langle d_k, Ax_k - b\rangle + \frac{1}{2}\langle d_k, Ad_k\rangle^2$. Thus $\frac{d}{d\lambda} J(x_k + \lambda d_k) = 0 \Leftrightarrow \lambda\langle d_k, Ad_k\rangle + \langle d_k, Ax_k - b\rangle = 0$. By strict convexity we obtain the result. □

Theorem 6.12. *Let $x_0 \in \mathbb{R}^n$ and $x_k)_k$ be defined by*

$$x_{k+1} = x_k + \lambda_k d_k \text{ with } \lambda_k = -\frac{\langle d_k, Ax_k - b\rangle}{\langle d_k, Ad_k\rangle}.$$

Then $(x_k)_k$ converges to x_* after at most n iterations with x_* such that $Ax_* - b = 0$.

Proof. The family $\{d_0, \cdots, d_{n-1}\}$ is a basis of $\mathbb{R}^n$. Thus there exists a unique $(\beta_0, \cdots, \beta_{n-1}) \in \mathbb{R}^n$ such that

$$x_* - x_0 = \sum_{i=0}^{n-1} \beta_i d_i \text{ with } \beta_i = \frac{\langle d_i, A(x_* - x_0)\rangle}{\langle d_i, Ad_i\rangle}.$$

But from the algorithm $x_k - x_0 = \sum_{i=0}^{k-1} \lambda_i d_i$. We deduce that $\langle d_k, A(x_k - x_0)\rangle = 0$. Hence

$$\begin{aligned}\langle d_k, A(x_* - x_0)\rangle &= \langle d_k, A(x_k - x_0)\rangle + \langle d_k, A(x_* - x_k)\rangle \\ &\langle d_k, b - Ax_k\rangle.\end{aligned}$$

We deduce that, for $0 \leq k \leq n-1$ we have $\beta_k = -\frac{\langle d_k, Ax_k - b\rangle}{\langle d_k, Ad_k\rangle} = \lambda_k$. thus after n iterations $x_n = x_* = x_0 + \sum_{i=0}^{n-1} \beta_i d_i = \sum_{i=0}^{n-1} \lambda_i d_i + x_0$. □

Algortihm 4:

- Choose $x_0 \in \mathbb{R}^n$; define $g_0 = Ax_0 - b$, $d_0 = -g_0$
- For $k \in \{0, \cdots, n-1\}$ if $g_k = 0$ stop; otherwise take
  - $\lambda_k = -\frac{\langle d_k, g_k\rangle}{\langle d_k, Ad_k\rangle}$
  - $x_{k+1} = x_k + \lambda_k d_k$
  - $g_{k+1} = Ax_{k+1} - b$
  - $\beta_{k+1} = \frac{\langle g_{k+1}, Ad_k\rangle}{\langle d_k, Ad_k\rangle}$
  - $d_{k+1} = -g_{k+1} + \beta_{k+1} d_k$.

The $(d_i)_{0 \leq i \leq n-1}$ are A-conjugated directions.

6.4. Algorithms for problems with constraints.

6.4.1. *The projected gradient method.* When dealing with a constraint set A which is closed and convex, we can use the projection $P : A \to \mathbb{R}^n$. Then a large class of descent methods with constraints can be dealt with with the following algorithm:

Algorithm:

- Choose $x_0 \in \mathbb{R}^n$.
- Define $x_{k+1} = P(x_k + \rho_k d_k)$ for some ρ_k, d_k constructed by one of the previous methods.

Theorem 6.13. *Let $A \subset \mathbb{R}^n$ be a closed convex set and $J : A \subset \mathbb{R}^n \to \mathbb{R}$ be an elliptic function (with ellipticity constant $\alpha > 0$). Assume that there exists $M > 0$ such that for every $x, y \in \mathbb{R}^n$*

$$\|\nabla J(x) - \nabla J(y)\| \leq M\|x - y\|.$$

If there exist $a, b > 0$ such that for all k, $0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2}$ then

- *the projected gradient method with variable step size ρ_k converges to the unique solution if the minimization problem x_*,*
- *the convergence is of order one $\|x_{k+1} - x_*\| \leq \beta\|x_k - x_*\|$ with*

$$\beta = \max\{\sqrt{1 - 2\alpha a + M^2a^2}, \sqrt{1 - 2\alpha b + M^2b^2}\}.$$

Proof. As x_* minimizes J in A, Euler's inequality implies that for all $y \in A$, $\langle \nabla J(x_*), y - x_*\rangle \geq 0$. Hence for all $y \in A$,

$$\langle x_* - \rho_k \nabla J(x_*) - x_*, y - x_*\rangle \leq 0.$$

This characterizes x_* as the projection on A of $x_* - \rho_k \nabla J(x_*)$, i.e.,

$$P(x_* - \rho_k \nabla J(x_*)) = x_*.$$

Consequently, following the proof in the case without constraints and using the fact that P is 1-Lipschitz (i.e., $\|P(x) - P(y)\| \leq \|x - y\|$) we find

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|P(x_k - \rho_k \nabla J(x_k)) - P(x_*)\|^2 \\ &\leq \|x_k - \rho_k \nabla J(x_k) - x_*\|^2 \\ &\leq (1 - 2\rho_k \alpha + \rho_k^2 M^2)\|x_k - x_*\|^2. \end{aligned}$$

Since $1 - 2\rho_k\alpha + \rho_k^2 M^2 < \beta^2 := \max\{1 - 2\alpha a + M^2 a^2, 1 - 2\alpha b + M^2 b^2\} < 1$ we find $\|x_{k+1} - x_*\|^2 \leq \beta \|x_k - x_*\|^2$. □

6.4.2. *Penalization methods.*

Theorem 6.14. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be a continuous, strictly convex, coercive function. Let A be a non-empty closed convex set of $\mathbb{R}^n$ and $\psi : \mathbb{R}^n \to \mathbb{R}$ be a continuous convex function such that $\psi(v) = 0$ for all $v \in \mathbb{R}^n$ and $\psi(v) = 0 \Leftrightarrow v \in A$. Then, for all $\epsilon > 0$, there exists a unique x_ϵ such that*

$$(P_\epsilon) \qquad x_\epsilon \in \mathbb{R}^n \text{ and } J_\epsilon(x_\epsilon) = \inf_{v \in \mathbb{R}^n} J_\epsilon(x) \text{ with } J_\epsilon(x) := J(x) + \frac{1}{\epsilon}\psi(x).$$

Moreover $\lim_{\epsilon \to 0} x_\epsilon = x_*$ where x_* is the unique solution of

$$\text{(P)} \qquad x_* \in A \quad \text{and} \quad J(x_*) = \inf_{x \in A} J(x).$$

Proof. (P) and (P_ϵ) have a unique solution (coercivity, continuity, strict convexity). Since

$$J(x_\epsilon) \leq J(x_\epsilon) + \frac{1}{\epsilon}\psi(x_\epsilon) = J_\epsilon(x_\epsilon) \leq J_\epsilon(x_*) = J(x_*),$$

we deduce that $(x_\epsilon)_\epsilon$ is bounded (by coercivity of J). Hence we can extract a subsequence $(x_{\epsilon'})'_\epsilon$ and find $\tilde{x} \in \mathbb{R}^n$ such that $\lim_{\epsilon' \to 0} x_{\epsilon'} = \tilde{x}$ (Bolzano-Weirstrass theorem). We use that $J(x_{\epsilon'}) \leq J(x_*)$ and the continuity of J to deduce

$$\lim_{\epsilon' \to 0} J(x_{\epsilon'}) = J(\tilde{x}) \leq J(x_*).$$

Since $0 \leq \psi(x_{\epsilon'}) \leq \epsilon'(J(x_*) - J(x_{\epsilon'}))$ and since $(x_{\epsilon'})_{\epsilon'}$ converges, $J(x_*) - J(x_{\epsilon'})$ is bounded. Hence $0 = \lim_{\epsilon' \to 0} \psi(x_{\epsilon'}) = \psi(\tilde{x})$. It means that $\tilde{x} \in A$. By uniqueness, we find $\tilde{x} = x_*$ and we also deduce that all the sequence converges toward x_* (since we can repeat the argument for every subsequences). □

Application: Convex programming Let $J : \mathbb{R}^n \to \mathbb{R}$ be strictly convex and $\varphi_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq m$, be convex functions. Find x_* such that $x_* \in A = \{x \in \mathbb{R}^n; \varphi(x) \leq 0, 1 \leq i \leq m\}$ and $J(x_*) = \inf_{x \in A} J(x)$. We can apply the penalization method to this problem with $\psi : x \in \mathbb{R}^n \mapsto \psi(x) = \sum_{i=1}^m \max\{\varphi_i(x), 0\}$.

Remark: It is often difficult to find good penalization functions ψ (although it is not a condition in the theorem we would like to have differentiability). Note that $\psi(x) = \sum_{i=1}^m \max\{\varphi_i(x), 0\}$ is not differentiable.

6.5. **Newton's method.** $V = \mathbb{R}^n$, this is a method designed to solve an equation of the form $F(x) = 0$. In optimization we will apply this method to $DJ(x) = 0$. We assume that F is C^2. Let x_* be a *regular* zero of F, i.e.,

$$F(x_*) = 0 \text{ and } DF(x_*) \text{ is invertible .}$$

We use Taylor-Young formula around x_* to write

$$0 = F(x_*) = F(x) + DF(x_*).(x - x_*) + o(\|x - x_*\|)$$

and thus

$$x_* = x - DF^{-1}(x_*).F(x) + o(\|x - x*\|).$$

The method consists in defining

$$\text{(N)} \qquad \begin{cases} \hfill x_0 \in \mathbb{R} \\ x_{n+1} = x_n - DF(x_n)^{-1}F(x_n) \text{ for } n \geq 0. \end{cases}$$

Theorem 6.15. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be of class C^2 and x_* be a regular zero of F (i.e., $F(x_*) = 0, DF(x_*) \in GL_n(\mathbb{R})$). There exists $\epsilon > 0$ such that if x_0 is such that $\|x_0 - x_*\| \leq \epsilon$ then the sequence defined by (N) converges to x_*. Moreover, there exists $C > 0$ such that*

$$\|x_{n+1} - x_*\| \leq C\|x_n - x_*\|^2.$$

Proof. By continuity of DF and since $GL_n(\mathbb{R})$ is open, there exists $\epsilon > 0$ such that $DF(x) \in GL_n(\mathbb{R})$ for every x such that $\|x - x_*\| < \epsilon$. Let us prove by induction that if $x_0 \in B(x_*, \epsilon)$ then $\|x_n - x_*\| < \epsilon$ for all n. The initialization is supposed to be true. Now if we assume that $\|x_n - x_*\| < \epsilon$, since $F(x_*) = 0$ we obtain, from the definition of x_{n+1}:

$$\begin{aligned} x_{n+1} - x_* &= x_n - x_* - DF(x_n)^{-1}(F(x_n) - F(x_*)) \\ &= x_n - x_* - DF(x_n)^{-1}.\left(DF(x_n).(x_n - x_*)\right) + DF(x_n)^{-1}.O(\|x_n - x_*\|^2) \\ &= DF(x_n)^{-1}.O(\|x_n - x_*\|^2). \end{aligned}$$

(By Taylor formula of order 2, here we use that $F \in C^2$). Since $\|x_n - x_*\| < \epsilon$ we deduce that there exists $C > 0$ (independent of n and related to the continuity modulus of DF, D^2F on $B(x_*, \epsilon)$ such that

$$\text{(6.5)} \qquad \|x_{n+1} - x_*\| \leq C\|x_n - x_*\|^2.$$

If ϵ is sufficiently small so that $C\epsilon \leq 1$ we deduce that $\|x_{n+1} - x_*\| \leq \epsilon$. Thus by induction $\|x_n - x_*\| \leq \epsilon$ for all n and we have (6.5). □

In optimization we apply Newton's method to the gradient ∇J. This gives

$$\text{(6.6)} \qquad \begin{cases} \hfill x_0 \in \mathbb{R}^n \\ x_{k+1} = x_k - (\text{Hess } J(x_k))^{-1}\nabla J(x_k). \end{cases}$$

There exist other generalized Newton methods (cf. [5])

Advantages:

- J is not necessarily elliptic nor convex.
- We have quadratic convergence.

Disadvantages:

- J has to be C^3 (actually Newton's method is still valid with $F \in C^1$ so J would have to be C^2).
- If x_0 is not close enough to x_* the method does not converge.

- there is a linear system to solve at each step.

7. Introduction to duality and Uzawa algortihm

7.1. **Introduction to duality.** Let V be a vector space of finite dimension and M be a set. We consider a function

$$L : V \times M \to \mathbb{R}.$$

Definition 7.1. We say that (x, λ) is a *saddle point* of L if

* x is a minimum for $L(\cdot, \lambda) : v \mapsto L(v, \lambda)$ and
* λ is a maximum for $L(x, \cdot) : \mu \mapsto L(x, \mu)$.

In other words (x, λ) is a saddle point of L if

$$\sup_{\mu \in M} L(x, \mu) = L(x, \lambda) = \inf_{v \in V} L(v, \lambda).$$

Theorem 7.2. *If (x, λ) is a saddle point of $L : V \times M \to \mathbb{R}$ then*

$$\sup_{\mu \in M} \inf_{v \in V} L(v, \mu) = L(x, \lambda) = \inf_{v \in V} \sup_{\mu \in M} L(v, \mu).$$

Proof. First we remark that we always have

$$\sup_{\mu \in M} \inf_{v \in V} L(v, \mu) \leq \inf_{v \in V} \sum_{\mu \in M} L(v, \mu). \tag{7.1}$$

Indeed if $\bar{v} \in V$ and $\bar{\mu} \in M$ we have

$$\inf_{v \in V} L(v, \bar{\mu}) \leq L(\bar{v}, \bar{\mu}) \leq \sup_{\mu \in M} L(\bar{v}, \mu).$$

Since $\inf_{v \in V} L(v, \bar{\mu})$ is a function of $\bar{\mu} \in M$ only and $\sup_{\mu \in M} L(\bar{v}, \mu)$ is a function of $\bar{v} \in V$ only we find that (7.1) is true. To prove the converse result we use that (x, λ) is a saddle point

$$\inf_{v \in V} \sup_{\mu \in M} L(v, \mu) \leq \sup_{\mu \in m} L(x, \mu) = L(x, \lambda) = \inf_{v \in V} L(v, \lambda) \leq \sup_{\mu \in M} \inf_{v \in V} L(v, \mu).$$

□

We suppose that $A = \{x \in \mathbb{R}^n ; \varphi(x) \leq, 1 \leq i \leq m\}$ for some functions $\varphi_i : \mathbb{R}^n \to \mathbb{R}$. We consider the problem

$$J(x_*) = \inf_{x \in A} J(x) \tag{$\mathcal{P}$}$$

We call Lagrangian associated to problem ($\mathcal{P}$) the function

$$\begin{array}{rlcl} L & : V \times \mathbb{R}^m_+ & \to & \mathbb{R} \\ & (x, \mu) & \mapsto & J(x) + \sum_{i=1}^m \mu_i \varphi_i(x). \end{array}$$

Theorem 7.3. i) *If $(x_*, \lambda) \in V \times \mathbb{R}^m_+$ is a saddle point of L then $x_* \in A$ and x_* is a solution to the constrained problem $(\mathcal{P})$.*

ii) *Conversely, if the functions φ_i, $1 \le i \le m$ are convex, differentiable at a point $x_* \in A$ and the constraints are qualified then, if x_* is a solution to $(\mathcal{P})$ there exists at least one $\lambda \in \mathbb{R}^m_+$ such that $(x_*, \lambda) \in V \times \mathbb{R}^m_+$ is a saddle point of L.*

Proof. i) Since $L(x_*, \mu) \le L(x_*, \mu)$ for any $\mu \in \mathbb{R}^m_+$, we deduce that $\sum_{i=1}^m (\mu_i - \lambda_i)\varphi_i(x_*) \le 0$ for all $\mu \in \mathbb{R}^m_+$. We let $\mu_i \to 0$ and we find that $\varphi_i(x_*) \le 0$ for $1 \le i \le m$. This means that $x_* \in A$. We can then take $\mu = 0$ and we obtain $\sum_{i=1}^m \lambda_i \varphi_i(x_*) \ge 0$. But $\lambda_i \ge 0$ and $\varphi_i(x_*) \le 0$, thus we deduce that

$$x_* \in A \text{ and } \sum_{i=1}^m \lambda_i \varphi_i(x_*) = 0. \tag{7.2}$$

We also use that $L(x_*, \lambda) \le L(x, \lambda)$ for all $x \in V$ to obtain

$$\begin{aligned} J(x_*) &\le J(x) + \sum_{i=1}^m \lambda_i \varphi_i(x) \text{ for all } x \in V \\ \Rightarrow J(x_*) &\le J(x) \text{ for all } x \in A. \end{aligned}$$

ii) We can apply KKT theorem. If x_* is a solution of $(\mathcal{P})$, there exists $\lambda \in \mathbb{R}^m_+$ such that

$$\sum_{i=1}^m \lambda_i \varphi_i(x_*) = 0 \text{ and } \nabla J(x_*) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(x_*) = 0.$$

The first equality means that

$$L(x_*, \mu) = J(x_*) + \sum_{i=1}^m \mu_i \varphi_i(x_*) \le J(x_*) = L(x_*, \lambda) \ \forall \mu \in \mathbb{R}^m_+.$$

The second equality is a sufficient condition of minimization for the convex function

$$L(\cdot, \lambda) : x \mapsto J(x) + \sum_{i=1}^m \lambda_i \varphi_i(x).$$

Thus we can say that

$$L(x_*, \lambda) \le L(x, v) \text{ for all } x \in V.$$

Hence (x_*, λ) is a saddle point of L.

□

We have proved that, under some assumptions, finding a solution x_* to problem $(\mathcal{P})$ is the same as finding the first arguments of saddle points of the Lagrangian $L(x, \mu) = J(x) + \sum_{i=1}^m \mu_i \varphi_i(x)$. If we knew one of the second argument λ of one of these saddle points then we could replace $(\mathcal{P})$ by a problem without constraints

$$x_\lambda \in V, \quad L(x_\lambda, \lambda) = \inf_{x \in V} L(x, \lambda). \tag{$\mathcal{P}_\lambda$}$$

How to find such a $\lambda \in \mathbb{R}^m_+$?

If we recall that $L(x_\lambda, \lambda) = \inf_{x \in V} L(x, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{x \in V} L(x, \mu)$ we can look for λ as the solution of

$$(\mathcal{Q}) \qquad \lambda \in \mathbb{R}^m_+ \quad G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu),$$

where $G : \mathbb{R}^m_+ \to \mathbb{R}$ is defined by

$$G(\mu) = \inf_{x \in V} L(x, \mu).$$

The problem ($\mathcal{Q}$) is called the *dual problem.* Note that ($\mathcal{P}$) is equivalent to

$$x_0 \in V \quad \mathcal{I}(x_) = \inf_{x \in V} \mathcal{I}(x) \text{ with } \mathcal{I}(x) = \sup_{\mu \in \mathbb{R}^m_+} L(x, \mu).$$

Indeed,

$$\sup_{\mu \in \mathbb{R}^m_+} J(x) + \sum_{i=1}^m \mu_i \varphi_i(x) = \begin{cases} J(x) \text{ if } x \in A \\ +\infty \text{ otherwise.} \end{cases}$$

Theorem 7.4. i) *We assume that $\varphi_i : V \to \mathbb{R}$, $1 \leq i \leq m$ are continuous and that for all $\mu \in \mathbb{R}^m_+$ the problem*

$$(\mathcal{P}_\mu) \qquad \textit{find } x_\mu \in V \textit{ such that } x_\mu \in V,\ L(x_\mu, \mu) = \inf_{x \in V} L(x, \mu)$$

has a unique solution which is a continuous function of $\mu \in \mathbb{R}^m_+$. Then, if λ is a solution to problem ($\mathcal{Q}$) then x_λ the solution to the corresponding problem ($\mathcal{P}_\lambda$) is a solution to ($\mathcal{P}$).

ii) *We assume that ($\mathcal{P}$) admits a solution x_*, we assume that $J, \varphi_i, 1 \leq i \leq m$ are convex and differentiable at x_*, if the constraints are qualified then ($\mathcal{Q}$) has at least one solution.*

Proof. i) Let λ be a solution to problem ($\mathcal{Q}$). We have $\lambda \in \mathbb{R}^m_+$ and $G(\lambda) = L(x_\lambda, \lambda) = \inf_{x \in V} L(x, \lambda)$. We will show that

$$\sup_{\mu \in \mathbb{R}^m_+} L(x_\lambda, \mu) = L(x_\lambda, \lambda).$$

This will prove that (x_λ, λ) is a saddle point of L and then by item i) of Theorem 7.3 we will deduce that x_λ is a solution to ($\mathcal{P}$). First we prove the differentiability of G. Let $\mu, \mu + \xi \in \mathbb{R}^m_+$

$$L(x_\mu, \mu) \leq L(x_{\mu+\xi}, \mu) \text{ and } L(x_{\mu+\xi}, \mu + \xi) \leq L(x_\mu, \mu + \xi)$$

this implies that

$$\sum_{i=1}^m \xi_i \varphi_i(x_{\mu+\xi}) \leq G(\mu + \xi) - G(\mu) \leq \sum_{i=1}^m \xi_i \varphi_i(x_\mu).$$

Thus, there exists $\theta \in [0, 1]$ such that

$$\begin{aligned} G(\mu + \xi) - G(\mu) &= (1 - \theta) \sum_{i=1}^m \xi_i \varphi(x_\mu) + \theta \sum_{i=1}^m \varphi_i(x_{\mu+\xi}) \\ &= \sum_{i=1}^m \xi_i \varphi_i(x_\mu) + \theta \sum_{i=1}^m \xi_i \left(\varphi_i(x_{\mu+\xi}) - \varphi_i(x_\mu)\right). \end{aligned}$$

The functions $\mu \in \mathbb{R}^m_+ \mapsto x_\mu \in V$ and $\varphi_i : V \to \mathbb{R}$ are continuous, thus

$$G(\mu+\xi) - G(\mu) = \sum_{i=1}^m \xi_i \varphi_i(x_\mu) + \|\xi\|\epsilon(\xi)$$

with $\lim_{\xi \to 0} \epsilon(\xi) = 0$. This means "almost"[7] that G is differentiable. Furthermore we have $DG(\mu).\xi = \sum_{i=1}^m \xi_i \varphi_i(x_\mu)$ for all $\xi \in \mathbb{R}^m$. The function G admits a maximum λ in $\mathbb{R}^m_+$ which is convex, thus from Proposition 4.2 we have

$$DG(\lambda).(\mu - \lambda) \le 0 \quad \forall \mu \in \mathbb{R}^m_+.$$

This means that $\sum_{i=1}^m \mu_i \varphi_i(x_\lambda) \le \sum_{i=1}^m \lambda_i \varphi_i(x_\lambda)$ for all $\mu \in \mathbb{R}^m_+$. Thus

$$\begin{aligned} L(x_\lambda, \mu) &= J(x_\lambda) + \sum_{i=1}^m \mu_i \varphi_i(x_\lambda) \\ &\le J(x_\lambda) + \sum_{i=1}^m \lambda_i \varphi_i(x_\lambda) = L(x_\lambda, \lambda) \quad \forall \mu \in \mathbb{R}^m_+. \end{aligned}$$

This means that (x_λ, λ) is a saddle point of L.

ii) From what precedes there exists $\lambda \in \mathbb{R}^m_+$ such that (x_λ, λ) is a saddle point of L. We apply the first Theorem 7.3 to obtain

$$L(x_*, \lambda) = \inf_{x \in V} L(x, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{x \in V} L(x, \mu)$$

and thus

$$G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{x \in V} L(x, \mu) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu).$$

□

7.2. **Uzawa's method.** Uzawa's algorithm is just the projected gradient method applied to the dual problem. The projection operator is simply computed as

$$P_+(\mu) = (\max(\mu_i, 0))_i \, .$$

Uzawa's algorithm:

- Start from $\mu_0 \in \mathbb{R}^m_+$.
- Given x_{k-1} and μ_k construct x_k and μ_{k+1} as follows
  - $->$ x_k is a minimizer of the unconstrained problem $\min_{x \in V} L(x, \mu_k) = G(\mu_k)$.
  - $->$ Choose $\rho_k (= \rho)$
  - $->$ Set $\mu_{k+1} = P_+ \left(\mu_k + \rho_k \nabla G(\mu_k)\right)$ with $\nabla G(\mu_k) = (\varphi_i(x_k))_i$ (seen previously)

Theorem 7.5. *We assume that J is elliptic and that A is of the form $A = \{x \in \mathbb{R}^n; Cx \le d\}$, $C \in M_{m,n}(\mathbb{R})$, $d \in \mathbb{R}^m$, $A \neq \emptyset$. Then, if $\rho_k = \rho$ and $0 < \rho < \frac{2\alpha}{\|C\|^2}$ then $(x_k)_k$ converges to the unique solution of ($\mathcal{P}$). If rank$(C) = m$ then λ_k converges to the unique solution of the dual problem ($\mathcal{Q}$).*

[7] The problem is that G is not defined in an open set, but what we proved is a kind of differentiability sufficient for our argument.

References

[1] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[2] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[3] H. Brezis and H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

[4] G. Carlier. *Classical and Modern Optimization*. World Scientific, 2022.

[5] P. G. Ciarlet, B. Miara, and J.-M. Thomas. *Introduction to numerical linear algebra and optimisation*. Cambridge university press, 1989.

Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.

Email address: luca.nenna@universite-paris-saclay.fr