Blanche Buet
blanche.buet@math.u-psud.fr

2017 – 2018
Orsay

A crash course in Optimization

Version provisoire, coquilles garanties ...

1. Some examples

1.1. **The shortest path and dynamic programming.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, and $w : \mathcal{E} \to \mathbb{R}_+$ a non negative weight function on $\mathcal{G}$. A path is a sequence of vertices $p = (v_1, \dots, v_n) \in \mathcal{V}^n$ such that two successive vertices v_i and v_{i+1} in p are connected by an edge in the graph, which is denoted $[v_i v_{i+1}]$. The length of such a path is then

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

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.

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 (think of 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.
- Geodesic distance when the metric is given as an image: fast marching algorithm (see the webpage of Gabriel Peyré on Dijkstra and Fast Marching Algorithms in his Numerical Tours).

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 such a problem, the key lies in the following observation: if p is a shortest path, then any subpath of p is also a shortest path, so that it is enough to consider path that are extensions 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 at the vertex.
(3) Iterate the process.

This is a rough description of *Dijkstra algorithm.*

1.2. **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 coastline) 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 determine, 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 rewrite

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

And the area to maximize is then $\int_a^b u(x)dx$. The answer is that Dido should enclose a half circle (a and b being adapted to the length L), but try to prove it...

The catenary's problem. A cable of fixed length and uniform mass density is suspended between two electric pylons, what is the shape of the chain, what must be the minimal height of pylons so as that the chain does not touch the ground? If we model the problem as previously, it amounts to minimize the potential energy of the chain

$$E(u) = \int_a^b u(x)\sqrt{1+u'(x)^2}dx$$

under the length constraint (1). The answer is not a parabola as Galileo claimed, but a catenary

$$u(x) = \alpha \cosh(x/\alpha)\,,\ \alpha > 0.$$

An image processing problem. 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), find u minimizing

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

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

1.3. **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), Toulouse (y_2) and Grenoble (y_3), they want to minimize to 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's rewrite the problem, if c_{ij} denotes the cost if x_i goes to the town y_j, then the tabular above is the matrix $(c_{ij})_{ij}$. Let $A = (c_{ij})_{ij}$ be defined as $a_{ij} = 1$ if x_i goes to y_j and 0 otherwise. We thus want to find A minimizing

$$\sum_{i,j=1}^{3} a_{ij}c_{ij}$$

under the constraints

$$\begin{array}{l} a_{ij} \in \{0,1\} \\ \sum_j a_{ij} = 1 \text{ person } i \text{ goes to exactly one town} \\ \sum_i a_{ij} = 1 \text{ town } j \text{ is reached by exactly one person} \end{array}$$

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

$$a_{ij} \in [0,1] \text{ that is } a_{ij} \leq 1 \text{ and } a_{ij} \geq 0\,.$$

without changing the minimizers and it amounts to minimize a *linear cost* under *linear constraints*. Of course, 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 while the *Hungarian algorithm* allows to solve it in $O(n^3)$.

1.4. **Study of an optimization problem.**

- **Existence** of minimizers?
- **Characterization** of minimizers?
- **Numerical computation** of a minimizer?

2. Existence

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: find $x_* \in A$ such that

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

The results of existence depend strongly on the fact that V is finite or infinite dimensional. The explanation being quite simple, existence is provided thanks to compactness and characterizing compact sets is quite different in finite or infinite dimension. Uniqueness results hold essentially under convexity assumptions.
We begin with recalling a few basic definitions.

2.1. **Some definitions and notations.**

Definition 1 (Global, local minimizer)**.** *An element $x_* \in A$ is a* global minimizer *of J on A iff*

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

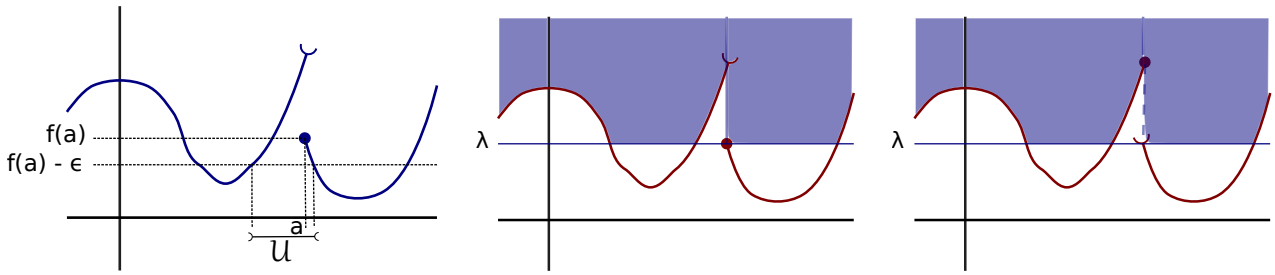
An element $x_* \in A$ is a local minimizer *of J on A iff*

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

Definition 2 (Minimizing sequences)**.** *A* minimizing sequence *of J in A is a sequence $(x_n)_n \subset A$ such that $J(x_n) \xrightarrow[n\to\infty]{} \inf_A J$.*

2.2. **Lower semi continuity.** We give definitions in the framework of topological spaces because it will be important to consider the *weak topology* on V, rather than the one induced by the norm, in order to increase the amount of compact sets.

Definition 3 (Lower semi continuity l.s.c., see Figure 1(a))**.** *Let X be a topological space and $f : X \to \mathbb{R} \cup \pm\infty$, then f is* lower semi continuous *(l.s.c.) at $a \in X$ iff for every $\epsilon > 0$ there exists a neighbourhood U of a such that $f(x) \geq f(a) - \epsilon$ for all $x \in U$ when $f(a) < +\infty$, and $f(x)$ tends to $+\infty$ as x tends to a when $f(a) = +\infty$.*



(a) (b) (c)

FIGURE 1

In orther 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 4 (Epigraph)**.** *Let X be a topological space and $f : X \to \mathbb{R} \cup \{\pm\infty\}$, then the* epigraph *of f is the set of points lying above its graph:*

$$\mathrm{epi}(f) = \{(x, \lambda) \in V \times \mathbb{R} \cup \{\pm\infty\} \,|\, \lambda \geq f(x)\} .$$

Proposition 1. *Let $f : X \to \mathbb{R} \cup \{\pm\infty\}$, then f is lower semi continuous iff* epi(f) *is closed (in $X \times \mathbb{R} \cup \{\pm\infty\}$ endowed with the product topology).*

In Figure 1(b), epi(f) is closed and f is lower semi continuous while in Figure 1(c), epi(f) is not closed and f is not lower semi continuous.

Definition 5 (Sequential lower semi continuity)**.** *Let X be a topological space and $f : X \to \mathbb{R} \cup \{\pm\infty\}$, then f is* lower semi continuous *(l.s.c.) at $a \in X$ iff for all sequence $(x_n)_n$ tending to a in X,*

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

Remark 1 (Continuity and 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 or more generally in first-countable topological spaces (each point has a countable neighbourhood basis). Moreover, in all the results of existence of minimizers stated in these notes, only the sequential lower semi continuity is needed.

2.3. **Direct method in the calculus of variations to prove existence.** Let τ be a topology on X.

(1) Take a minimizing sequence $(x_n)_n \subset A$ and show that it admits a subsequence $(x_{n_k})_k$ converging for τ to some $x_* \in A$. This is a compactness issue.
(2) Show that J is sequentially lower semi continuous with respect to the topology τ on X.
(3) In this case,
$$\inf\{J(x) \,|\, x \in A\} = \lim_{n\to\infty} J(x_n) = \lim_{k\to\infty} J(x_{n_k}) \geq J(x_*)\,.$$

Theorem 2 (Existence)**.** *Assume that $J : A \to \mathbb{R} \cup \{+\infty\}$ is lower semi continuous and that A is a non empty compact set, assume moreover that J is not constantly equal to $+\infty$ on A, then there exists at least one minimizer of J in A.*

Proof. Let $(x_n)_n \subset A$ be a minimizing sequence for J i.e. $\lim_n J(x_n) = \inf_A J$. By compactness of A, there exists a subsequence $(x_{n_k})_k$ converging to some $x_* \in A$. By l.s.c. of J, we have
$$J(x_*) \leq \liminf_k J(x_{n_k}) = \lim_n J(x_n) = \inf_A J.$$
□

Remark 2. Proof of Theorem 2 is valid for any topology (for which J is l.s.c. and A is compact).

In finite dimension, compact sets are easy to characterize, they are closed bounded sets, which is not true in infinite dimension.

2.4. **Existence in finite dimension.**

Theorem 3 (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.*

Definition 6 (Coercivity)**.** *Let V be a normed vector space. A function $f : A \subset A \to \mathbb{R}$ is* coercive *(or* infinite at infinity*) iff*
$$\lim_{\substack{\|x\|\to+\infty \\ x\in A}} f(x) = +\infty\,.$$

2.5. **The case of a quadratic functional.** This is a very important particular case. Let $V, <\cdot,\cdot>$ be a **Hilbert** space and $J : V \to \mathbb{R}$ be defined as
$$J(v) = \frac{1}{2}a(x,x) - b(x)\,, \tag{2}$$
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, the existence of a minimizer is easier to prove, it follows from the projection theorem and Riesz representation theorem.

Theorem 4. *Let $J : V \to \mathbb{R}$ be a quadratic functional defined on the Hilbert space V as in* (2)*. Assume that the bilinear form a is elliptic i.e. there exists $\alpha > 0$ such that for all $x \in V$,*
$$a(x,x) \geq \alpha\|x\|^2\,.$$
Assume 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
$$a(x_*, x - x_*) \geq b(x - x_*) \text{ for all } x \in A\,. \tag{3}$$

Proof. • As a is elliptic, it defines another scalar product on V and 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 thus it is equivalent to minimize $a(x-v, x-v) = \|x-v\|_a^2$ for $x \in A$, if $\|\cdot\|_a$ 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 3. In the case $A = V$, characterization (3) simply rewrites: for all $x \in V$,
$$a(x_*, x) = b(x)\,.$$

Example 1 (Least square approximation). Let $A \in M_{n,p}$ and $B \in M_{n,1}$ with $n > p$. The linear system $Ax = B$ is then generally overdetermined and we consider the solution in the sense of least square approximation, that is $x_* \in \mathbb{R}^p$ minimizing
$$J(x) = \|Ax - B\|^2 = \frac{1}{2}\left\langle A^TAx, x\right\rangle - \left\langle A^TB, x\right\rangle + \frac{1}{2}\|B\|^2\,.$$
Characterization (3) then give the *normal equation*:
$$A^TAx_* = A^TB\,,$$
and the matrix A^TA is positive definite if A has rank p.
In the case of a polynomial fitting (of order 3 for instance) of n points $\{(x_i, y_i)\}_{i=1\dots n} \subset \mathbb{R}^2$, for all $i = 1 \dots n$, $y_i = \alpha + \beta x_i + \gamma x_i^2 + \delta x_i^3$ and
$$A = \begin{pmatrix} 1 & x_1 & (x_1)^2 & (x_1)^3 \\ \vdots & & & \vdots \\ 1 & x_n & (x_n)^2 & (x_n)^3 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix}$$

Example 2 (Variational formulation of elliptic problem). Let $\Omega \subset \mathbb{R}^n$ be a smooth bounded[1] open set and $f \in \mathrm{L}^2(\Omega)$, let $J : \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ 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 $H_0^1(\Omega)$ and moreover u is solution of the variational formulation in $\mathrm{H}_0^1(\Omega)$ of the Laplacian equation
$$-\Delta u = f \text{ in } \Omega\,.$$
Indeed, let a and b be the applications defined as
$$\begin{array}{rcccl} a & : & \mathrm{H}_0^1(\Omega)\times\mathrm{H}_0^1(\Omega) & \to & \mathbb{R} \\ & & (u,v) & \mapsto & \int_\Omega \langle \nabla u, \nabla v\rangle \end{array} \quad \text{and} \quad \begin{array}{rcccl} b & : & \mathrm{H}_0^1(\Omega) & \to & \mathbb{R}. \\ & & u & \mapsto & \int_\Omega fu \end{array}$$
- The linear form b is continuous: for every $u \in \mathrm{H}_0^1(\Omega)$, $|b(u)| \leq \|f\|_{\mathrm{L}^2}\|u\|_{\mathrm{L}^2} \leq \|f\|_{\mathrm{L}^2}\|u\|_{\mathrm{H}_0^1}$.
- The symmetric bilinear form a is continuous: for every $u, v \in \mathrm{H}_0^1(\Omega)$, $|a(u,v)| \leq \|\nabla u\|_{\mathrm{L}^2}\|\nabla v\|_{\mathrm{L}^2} \leq \|u\|_{\mathrm{H}_0^1}\|v\|_{\mathrm{H}_0^1}$.
- a is elliptic: thanks to Poincaré's inequality, there exists a constant C, only depending on Ω and such that $\|u\|_{\mathrm{H}_0^1} \leq C\|\nabla u\|_{\mathrm{L}^2}$. Therefore, for $u \in \mathrm{H}_0^1(\Omega)$,
$$a(u,u) \geq \frac{1}{2C^2}\|u\|_{\mathrm{H}_0^1}^2.$$

It remains to apply Theorem 4 to obtain a unique minimizer $u \in \mathrm{H}_0^1(\Omega)$ satisfying for all $v \in \mathrm{H}_0^1(\Omega)$,
$$\int_\Omega \langle \nabla u, \nabla v\rangle = \int_\Omega fv.$$

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

As we already mentioned, the flaw in infinite dimension, is that in general, being compact is stronger than being closed and bounded. When the topology comes from a norm, Riesz Theorem (the unit ball of a Banach space is compact iff it has finite dimension) even says that it is always false. Yet, there exists a well-known topology which makes the closed unit ball compact: the *weak–∗ topology.*

2.6. **Weak topology and compact sets.** Loosely speaking, the idea is the following, let (X, τ) be a topological space:

- the less open sets in the topology τ, the more there are compact sets w.r.t. τ,
- the more open sets in the topology τ, the more there are continuous functions $f : (X, \tau) \to \mathbb{R}$.

Hence, while changing the topology to try to make closed bounded sets compact, we may lose the continuity or lower semi continuity of J.

Definition 7 (Weak topology)**.** *The* weak topology *on V is the smallest (=least fine= weakest) topology (i.e. containing the least possible open sets) such that all $L \in V'$ are continuous.*

In other words, the weak topology is the topology generated by $\{L^{-1}(U) \,|\, L \in V' \text{ and } U \subset \mathbb{R} \text{ open}\}$[2]. In finite dimension, the weak topology and the strong topology (induced by the norm $\|\cdot\|$) coincide. In infinite dimension, the weak topology is strictly smaller than the strong topology.

Proposition 5 (Weak convergent sequences)**.** *Let V be a Banach space and $(x_n)_{n\in\mathbb{N}} \subset V$. Then,*

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

Pay attention that when V has infinite dimension, the weak topology behaves very differently from the norm topology. While the following implication are correct, the converse are completly false in general.
If a set is weakly open (resp. weakly closed) then it is open (resp. closed) for the norm topology.
If $f : V \to \mathbb{R}$ is weakly lower semicontinuous i.e. with respect to the weak topology on V, then it is lower semi continuous with respect to the norm topology on V.
For instance, it is possible to show that any weak neighbourhood of 0 contains a vector sub-space of infinite dimension.

Exercise 1. 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 will use. 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: convex sets. This is a consequence of Hahn-Banach Theorem.

Proposition 6. *Let $C \subset V$ be a convex set, then C is strongly closed iff C is weakly closed.*

And let us finally state the result we were interested in, that is weak compactness of convex closed bounded sets.

Definition 8 (Reflexive Banach space)**.** *A reflexive Banach space is a Banach space V isomorphic*[3] *to its bi-dual V'' via the canonical injection,*

$$\begin{array}{rccl} J : & V & \to & V'' \\ & x & \mapsto & \left(\begin{array}{rccl} \Phi_x : & V' & \to & \mathbb{R} \\ & L & \mapsto & L(x) \end{array} \right) \end{array}$$

Remark 4. It is easy to check that the linear injection is continuous and that $\|J(x)\|_{V''} \leq \|x\|$. Indeed,

$$\|J(x)\|_{V''} = \sup_{\|\zeta\|_{V'}=1} |\Phi_x(\zeta)| = \sup_{\|\zeta\|_{V'}=1} |\zeta(x)| \leq \|\zeta\|_{V'}\|x\| \leq \|x\| \,.$$

Proposition 7. *Let V be a reflexive Banach space, then the canonical injection J is an isometry: for all $x \in V$,*

$$\|J(x)\|_{V''} = \|x\| \,.$$

[2]Any intersection of topologies is a topology, thus the topology generated by a set S is simply the intersection of all topologies containing S.

[3]The vector space isomorphism automatically implies that J is bi–continuous and even that J is an isometry.

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 $L(y) = t$ where $y = t\frac{x}{\|x\|} \in \mathbb{R}x$. 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, $\zeta_0 \in V'$, such that $\|\zeta_0\|_{V'} = \|L\|_{(\mathbb{R}x)'} = 1$ and $\zeta_0 = L$ in $\mathbb{R}x$. In particular, $\zeta_0(x) = L(x) = \|x\|$. We can thus conclude that

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

□

Proposition 8. *Let V be a reflexive Banach space and $(x_n)_n \subset V$ a sequence weakly converging to x. Then,*

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

Proof.

- It is a general property of weak topology in a reflexive Banach space $(V, \|\cdot\|)$, the Banach norm is l.s.c. with respect to the associated weak topology. It is a consequence of the (strong) continuity and convexity of the norm.
- It is a consequence of Banach–Steihauss theorem and the fact that J is an isometry. Indeed, let us consider the evaluation maps $\Phi_{x_n} = J(x_n)$ and $\Phi_x = J(x)$. As x_n weakly converges to x, for any $\zeta \in V'$, $\zeta(x_n) \xrightarrow[n\to\infty]{} \zeta(x)$ that is, $\Phi_{x_n}(\zeta) \xrightarrow[n\to\infty]{} \Phi_x(\zeta)$. Hence, the sequence of continuous linear forms (Φ_{x_n}) pointwise converges to Φ_x. Banach–Steihauss theorem implies that $\sup_n \|\Phi_{x_n}\|_{V''} < +\infty$ and $\|\Phi_x\|_{V''} \leq \liminf_{n\infty} \|\Phi_{x_n}\|_{V''}$. We conclude with the fact that $\|\Phi_{x_n}\|_{V''} = \|x_n\|$ and $\|\Phi_x\|_{V''} = \|x\|$ since J is an isometry.

□

Theorem 9. *Let V be a **reflexive** Banach space. Then convex bounded closed sets are weakly compact.*

Remark 5 (Weak–$*$ topology). It is also possible to define the weak topology on V', and moreover, it is possible to weaken even more the weak topology, reducing the set of continuous applications: instead of considering the topology making all elements of $(V')'$ continuous, we define the topology making continuous all the elements of

$$\left\{ \begin{array}{ccc} V' & \to & \mathbb{R} \\ L & \mapsto & L(x) \end{array} : x \in V \right\} \subset (V')' .$$

This is the weak–$*$ topology on V', and the weak–$*$ topology is smaller than the weak topology in V'. And for this topology, the well-known Banach-Alaoglu Theorem states that the closed unit ball of V' is compact.

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

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

We consider the problem of minimizing $\mathcal{E}$ in $\mathrm{H}^1(\Omega)$.

- **Coercivity**. As we have

  $$\mathcal{E}(u) \geq \|\nabla u\|^2_{\mathrm{L}^2} + \int_\Omega (u^2-1)^2 ,$$

  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,

  $$\int_\Omega (u^2-1)^2 \geq 2\|u\|^2_{\mathrm{L}^2} - 3|\Omega| ,$$

  which leads to $\mathcal{E}(u) \geq 2\|u\|^2_{\mathrm{H}^1} - 3|\Omega| \to +\infty$ when $\|u\|_{\mathrm{H}^1} \to +\infty$.
- **Compactness**. Notice that $\mathcal{E}$ is proper (i.e. $\mathcal{E}$ is not constantly equal to $+\infty$, actually, in our case, $\mathcal{E}(u) < +\infty$ for all $u \in \mathrm{H}^1$) and let u_n be a minimizing sequence, that is $\mathcal{E}(u_n) \xrightarrow{n\to\infty} \inf_{u\in\mathrm{H}^1} \mathcal{E}(u)$. From the coercivity of $\mathcal{E}$, we know that $(u_n)_n$ is bounded in H^1 and thus weakly compact in H^1.

- **Lower semi continuity.** Let $(u_n)_n$ be a sequence weakly converging to u in H^1. We want to prove that $\mathcal{E}(u) \leq \liminf_{n\infty} \mathcal{E}(u_n)$. Let us rewrite $\mathcal{E}$ as

$$\mathcal{E}(u) = \|u\|^2_{\mathrm{H}^1} - 2\int_\Omega fu + \|f\|_{\mathrm{L}^2} + \int_\Omega (u^2-1)^2.$$

As the norm is l.s.c with respect to the weak convergence 8 and the application $u \mapsto \int_\Omega fu$ is linear continuous in H^1 and thus H^1–weakly continuous (by definition of weak topology), it remains to prove that

$$\int_\Omega (u^2-1)^2 \leq \liminf_{n\infty} \int_\Omega ((u_n)^2-1)^2$$

Up to extraction, we can assume that

$$\liminf_{n\infty} \int_\Omega ((u_n)^2-1)^2 = \lim_{n\infty} \int_\Omega ((u_n)^2-1)^2.$$

As $(u_n)_n$ is H^1–weakly converging, it is bounded in H^1 (see Proposition 8, and since the canonical injection of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^2(\Omega)$ is compact[4], there is a subsequence $(u_{n_k})_k$ which converges strongly in L^2 to u.[5] We can thus extract again a subsequence $(u_{n_{k_l}})_l$ almost anywhere converging to u and by Fatou's Lemma

$$\int_\Omega (u^2-1)^2 \leq \liminf_{l\infty} \int_\Omega (u^2_{n_{k_l}}-1)^2 = \lim_{n\infty} \int_\Omega ((u_n)^2-1)^2.$$

And a sum of two l.s.c. functions is l.s.c. hence $\mathcal{E}$ is H^1–weakly l.s.c.
- **Conclusion.** We conclude with the direct method of calculus of variations. Let $(u_n)_n$ be a minimizing sequence of $\mathcal{E}$ in $\mathrm{H}^1(\Omega)$, as we showed its compactness, let $(u_{n_k})_k$ be a subsequence converging to $u_* \in \mathrm{H}^1(\Omega)$. As J is $\mathrm{H}^1(\Omega)$–weakly lower semi continuous, we have

$$\inf_{\mathrm{H}^1(\Omega)} J = \lim_{n\to\infty} J(u_n) = \liminf_{k\to\infty} J(u_{n_k}) \geq J(u_*)\,.$$

And u_* is a minimizer of J in $\mathrm{H}^1(\Omega)$.

Exercise 2. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and consider

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

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

However, in the previous example we, showed that $u \mapsto \int_\Omega (u^2-1)^2$ was weakly lower semi continuous while $f(t) = (t^2-1)^2$ is not convex in $\mathbb{R}$. Isn't there a contradiction ? Fortunately no, the subtlety lies in the fact that there are as many weak topologies as there are norms. And $\mathcal{E}$ is not L^2–weakly lower semi continuous but H^1–weakly lower semi continuous, which is weaker (and actually, we used strong L^2–convergence through compact injection of H^1 into L^2.

2.8. **Existence in infinite dimension.**

Exercise 3. Check that $J : A \to \mathbb{R}$ is convex $\Leftrightarrow$ epi(J) is convex.

Theorem 10. *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 semicontinuous and coercive, then there exists a minimizer of J in A.*

Proof.
- J convex $\Rightarrow$ epi(J) convex therefore: J is lower semicontinuous iff epi(J) is strongly closed iff epi(J) is weakly closed iff J is lower semicontinuous for the weak topology.
- J is coercive $\Rightarrow J(x) \geq M$ for all x such that $\|x\| \geq R$.
- Let B be the closed ball of radius R then, $A \cap B$ is a closed convex set $\Rightarrow A \cap B$ is a weak closed convex set $\Rightarrow A \cap B$ is weakly compact.

[4] Pay attention to the assumptions on Ω, the injection of $\mathrm{H}^1(\mathbb{R}^n)$ into $\mathrm{L}^2(\mathbb{R}^n)$ is not compact.

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

- J is weakly lower semicontinuous on a weakly compact set $\Rightarrow$ there exists $x_* \in A \cap B$ a minimizer of J on $A \cap B$ and for all $x \in A \setminus B$, $J(x) \geq M \geq J(x_*)$.

□

Finally, let us give an example without the convexity assumption and where the existence of a minimizer fails.

Exercise 4. We consider the Hilbert (thus reflexive) space

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

provided with the scalar product $(x_n)_n \cdot (y_n)_n = \sum_{n=0}^{\infty} x_n y_n$, and we define

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

Check that f is coercive and lower semicontinuous and check that however, f does not admit minimizer on $l^2(\mathbb{R})$.

3. Optimality conditions

3.1. **A bit of calculus...** The aim of this section is only to freshen up fundamental definitions and properties in calculus, which we will use in the sequel.

3.1.1. *Differentiability.*

Definition 9. *Let $(V, \|\cdot\|_V), (W, \|\cdot\|_W)$ be two normed vector spaces and $f : \Omega \subset V \to W$ be defined in 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 6. In infinite dimension, the differentiability depends on the norms on E and F. In finite dimension, as all the norms are equivalent, the differentiability does not depend on the choice of norms.

The application f is said to be C^1 if f is differentiable at every point of Ω and the application

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

is continuous.

Definition 10 (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 7. If f is differentiable at x_0 then f admits directional derivative in any direction $h \in V$ at x_0 and

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

it 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.

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

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

It is called the *gradient of* f. If $V = \mathbb{R}^n$ is endowed with the usual Euclidean scalar product and $(e_1, \dots, e_n)$ 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 5. 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 its hessian.

Correction 1. f is C^∞ by composition and for $x = (x_1, x_2, x_3) \in (\mathbb{R}_+^*)^3$ and $h = (h_1, h_2, h_3) \in \mathbb{R}^3$,

$$Df(x)(h) = < \nabla f(x), h > = \partial_1 f(x) h_1 + \partial_2 f(x) h_2 + \partial_3 f(x) h_3 \tag{4}$$

$$= (1 + \ln x_1) h_1 + (1 + \ln x_2) h_2 + (1 + \ln x_3) h_3 \tag{5}$$

and

$$\nabla^2 f(x) = \begin{pmatrix} 1/x_1 & 0 & 0 \\ 0 & 1/x_2 & 0 \\ 0 & 0 & 1/x_3 \end{pmatrix} .$$

Exercise 6 (An example in infinite dimension). Let $E_0 = \{u \in \mathrm{C}^1([a,b], \mathbb{R})\, u(a) = u(b) = 0\}$ be the vector space normed with $\|u\|_{\mathrm{C}^1} := \sup_{[a,b]} |u| + |u'|$. 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.

Correction 2. Let $u, h \in E_0$,

$$\begin{aligned} J(u+h) - J(u) &= \int_a^b \big[L(x, u(x) + h(x), u'(x) + h'(x)) - L(x, u(x) + h(x), u'(x) + h'(x)) \big] \\ &= \int_a^b < \nabla L(x, u(x), u'(x)), (0, h(x), h'(x)) > + o((0, h(x), h'(x)))\, dx \\ &= l_u(h) + r_u(h) , \end{aligned}$$

where for a fixed $u \in E_0$,

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

and $r_u(h) = J(u+h) - J(u) - l_u(h)$. The application l_u is

- linear,
- continuous since, applying triangular inequality and Cauchy Schwartz, we have for all $h \in E_0$,

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

It remains to prove that $r_u(h) = o(\|h\|_{\mathrm{C}^1})$. First step, let us prove that for all $K \subset \mathbb{R}^3$ compact set, for all $\epsilon > 0$,

$$\exists \eta > 0,\, \forall X \in K,\, \forall H \in \mathbb{R}^3, \quad \|H\| < \eta \Rightarrow |L(X+H) - L(X) - < \nabla L(X), H > | \le \epsilon \|H\| .$$

Let $K' = \{X + H \in \mathbb{R}^3 : X \in K, \|H\| \le 1\}$ 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 $\nabla^2 L$ 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 $\forall X, Y \in K'$,

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

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

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

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

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

Second step, apply first step for all $x \in [a,b]$ with $X(x) = (x, u(x), u'(x))$ and $H(x) = (0, h(x), h'(x))$. 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 $K = B_R$ of center 0 and radius R. Moreover, for all $x \in [a,b]$,

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

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

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

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

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

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

is differentiable at x_0. The resulting differential

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

is identified with the continuous bilinear application

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

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

Theorem 11 (Taylor's formulas)**.** *Let $f : \Omega \subset V \to W$ be defined in an open space Ω of a vector space V and $x_0 \in \Omega$. If f is differentiable at x_0, then*

$$f(x_0 + h) = f(x_0) + Df(x_0)(h) + o(\|h\|) \ .$$

If f is twice differentiable at x_0, then

$$f(x_0 + h) = f(x_0) + Df(x_0)(h) + \frac{1}{2} D^2 f(x_0)(h,h) + o(\|h\|^2) \ .$$

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

$$f(x_0 + h) = f(x_0) + Df(x_0)(h) + \frac{1}{2} D^2 f(x_0 + \theta h)(h,h) \ .$$

Remark 8. Taylor's formula of order 1 is just the differentiability at x_0, while Taylor'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

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

where $\nabla^2 f(x_0)$ is the *hessian* of f at x_0, that is the symmetric matrix defined as

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

3.2. **Necessary conditions.** In this section and the following one, the minimization problem is set in an **open** set A.

Proposition 12. *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: Euler's equation

$$DJ(x_*) = 0 \,.$$

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

$$D^2 J(x_*)(h, h) \geq 0 \,. \tag{6}$$

Remark 9. If V is finite dimensional, (6) exactly means that the Hessian $\nabla^2 J(x_*)$ is semi positive 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 $\phi'(0) = 0 = DJ(x_*)(h)$. And 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

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

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 still have the following result, giving a necessary condition in directions in which it is possible to make small variations from x_*.

Proposition 13. *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 $[x_*, x_* + h] \subset A$.

Exercise 7. Prove the previous proposition.

Correction 3. 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 13 is known as *Euler inequality* and rewrites:

Proposition 14 (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 13 to the direction $x - x_* \in V$. □

3.3. **Sufficient conditions.** We now look for conditions insuring that some 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 (6) is not sufficient as well (0 satisfies this condition for $x \mapsto x^5$).

Theorem 15. *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 $h \in V$,*

$$D^2J(x_*)(h,h) \geq \alpha \|h\|^2 \,, \tag{7}$$

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

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

  then x_* is a relative minimum for J.

Remark 10. Notice that if V is finite dimensional, condition (7) is equivalent to say that the Hessian $\nabla^2 J(x_*)$ is positive definite.

Proof. Follows from Taylor-Young and Taylor-Maclaurin formulas. □

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

Correction 4.

- *Existence:* f is continuous. As $2xy \leq x^2 + y^2$,

$$f(x,y) \geq x^4 + y^4 - 2x^2 - 2y^2 = (x^2-1)^2 - 1 + (y^2-1)^2 - 1 \xrightarrow[\|(x,y)\| \to 0]{} +\infty$$

  therefore f is coercive and there exists a minimizer.
- *Critical points:* f is C^1 and $\nabla f(x,y) = (4x^3 - 4y, 4y^3 - 4x)$ so that

$$\nabla f(x,y) = 0 \quad \Leftrightarrow \quad \left\{ \begin{array}{ll} y & = x^3 \\ x = x^9 & \end{array} \right. \quad \Leftrightarrow \quad \left\{ \begin{array}{ll} y & = x^3 \\ x & \in \{0,1,-1\} \end{array} \right.$$

  and critical points are $(0,0)$, $(-1,-1)$ and $(1,1)$.
- *Hessian:*

$$\nabla^2 f(x,y) = \begin{pmatrix} 12x^2 & -4 \\ -4 & 12y^2 \end{pmatrix}$$

  At $(0,0)$, $\det \nabla^2 f(0,0) < 0$ so that $(0,0)$ is a saddle point. At $(1,1)$ and $(-1,-1)$, $\det \nabla^2 f(1,1) > 0$ and $\mathrm{tr} \nabla^2 f(1,1) > 0$ so that $\nabla^2 f(1,1)$ is positive definite and $(1,1)$ and $(-1,-1)$ are local minima. Moreover

$$f(x,y) - f(1,1) = f(x,y) + 2 \geq (x^2-1)^2 + (y^2-1)^2 \geq 0 \,,$$

  and thus $(1,1)$ and $(-1,-1)$ are global minimizer. It was also possible to conclude given the existence of the minimizer and the fact that $f(-1,-1) = f(1,1)$.

Unfortunately, as well as the necessary conditions stated in the previous sections were not sufficient, these sufficient conditions are not necessary. Nevertheless, there is a particular but very important case where being a critical point is both necessary and sufficient to be a relative extremum: when the function J is convex.

3.4. **The case of convex functions.** Characterization of convexity of 1st and 2nd order and CNS of global minimum for convex functions (Cf. Ciarlet 7.4).

Proposition 16 (Characterization of convex functions)**.** *Let $A \subset V$ be an open set and $J : A \to \mathbb{R}$ be differentiable. Then*

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

If moreover J is twice differentiable, then

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

Order 1 characterization means that the graph of J is (strictly) above its tangent hyperplane everywhere.

Theorem 17. *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) the minimizers of J are exactly the critical points of J that is, 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 strict.*

4. Minimization with constraints

We now focus on the case where A is not open, we will consider constraints of the form

- *equality constraints*
- *inequality constraints*

4.1. Equality constraints.

In this section, we assume that

$$A = \{x \in V \,:\, g_1(x) = \ldots = g_p(x) = 0\}\,.$$

The solution to the minimization problem with equality constraints relies on the implicit functions theorem and simple linear algebra. We state the theorem in the general case where V is a Banach space and we will prove it in the particular case where V is finite dimensional: the geometric intuition is very helpful in this case.

Theorem 18 (Extrema liés)**.** *Let V be a Banach space and $J, g_1, \ldots, g_p : V \to \mathbb{R}$. 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*

(8) *the vectors $Dg_1(x_*), \ldots, Dg_p(x_*)$ are linearly independent.*

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

$$DJ(x_*) + \sum_{i=1}^{p} \lambda_i Dg_i(x_*) = 0\,. \tag{9}$$

Remark 11. If in addition we assume that J and g_i are convex, then the necessary condition of optimality given by Theorem 18 is also sufficient.

We begin with a glance at the linear case, where J is a continuous linear form in V and where the constraints are actually an intersection of hyperplanes. In this case, g_i are continuous linear form so that they are C^1 in V and for all $x \in V$, $Dg_i(x) = g_i$ and $DJ(x) = J$. Therefore, in this case (9) simply rephrases as

$$J + \sum_{i=1}^{p} \lambda_i g_i = 0\,. \tag{10}$$

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,

$$\forall h \in \bigcap_{i=1}^{p} \ker g_i,\ DJ(x)(h) = J(h) = 0 \quad \Leftrightarrow \quad \bigcap_{i=1}^{p} \ker g_i \subset \ker J\,. \tag{11}$$

Hence, the linear case only amounts to prove that (11) implies (10). This is a simple algebraic result which is an ingredient of the proof of Theorem 18.

Proposition 19 (the linear case: an algebraic result)**.** *Let $J, g_1, \ldots, g_p : V \to \mathbb{R}$ be linear forms. Assume that*

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

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

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

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

As $a \notin H$, $\lambda_0 = \langle \Lambda, a \rangle \neq 0$. □

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 esay 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 (10) and conclude with Proposition 19. We then give a less geometric proof in the general case of V being a Banach space and relying on the inverse function theorem.

Proposition 20 (Minimization over a submanifold)**.** *Let $M \subset V$ be a submanifold of V of dimension d. Let U be a neighbourhood of M, $J : U \to \mathbb{R}$ be differentiable and let x_* be a relative extremum of J on M, then $DJ(x_*)_{|T_{x_*}M} = 0$ (i.e. $\nabla J(x_*) \perp T_{x_*}M$).*

Proof. Let $h \in T_{x_*}M$, then 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 x_*, hence $(J \circ \gamma)'(0) = 0$ i.e. $DJ(\gamma(0))(\gamma'(0)) = \langle \nabla J(x_*), h \rangle = 0$. □

Proof. **Proof of Theorem 18 in finite dimension** The assumptions of Theorem 18 exactly mean that the application $g(x) = (g_1(x), \dots, g_p(x))$ is a submersion from a neighbourhood B of x_* onto $\mathbb{R}^p$ and thus

$$M = B \cap A = B \cap g^{-1}(0)$$

is a submanifold whose tangent plane at x_* is

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

Applying Proposition 20 leads to $\bigcap_{i=1}^{p} \ker Dg_i(x_*) \subset \ker DJ(x_*)$ and we conclude with Proposition 19 that $DJ(x_*) \in \operatorname{span}(\mathrm{Dg_1(x_*)}, \dots, \mathrm{Dg_p(x_*)})$. □

Proof. **Proof of Theorem 18 in a Banach space**
We start with recalling the inverse function theorem in Banach spaces: if V, W are Banach spaces, $\Omega \subset V$ is an open set and $l : \Omega \to W$ is 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 C^1–diffeomorphism when restricted from $\mathcal{V}$ to $\mathcal{W}$.

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 centered at 0 and such that

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

Define $F : \mathcal{U} \to \mathbb{R}^{p+1}$ s.t. for $x \in \mathcal{U}$, $F(x) = (J(x), g_1(x), \dots, g_p(x))$. Note that for all $c < J(0)$, $(c, 0, \dots, 0) \notin F(\mathcal{U})$ but $F(0) = (J(0), 0, \dots, 0) \in F(\mathcal{U})$, whence $F(\mathcal{U})$ cannot contain any open set around $F(0)$.
If $DF(0) \in \mathcal{L}(V, \mathbb{R}^{p+1})$ were surjective, then $G = \ker DF(0)$ 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 $\operatorname{Im} DF(0) = \mathbb{R}^{p+1}$ and $DF(0)_{|E} \in \operatorname{Isom}(E, \mathbb{R}^{p+1})$ (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 equipp E with the norm $\| \cdot \|_E$ induced by V (i.e. for all $x \in E$, $\|x\|_E = \|x\|_V$). Then $\tilde{F}$ is differentiable and

[6]Define $\hat{\Omega} = \Omega - x_*$, $\hat{J} : \hat{\Omega} \to \mathbb{R}^{p+1}$, $\hat{J}(x) = J(x + x_*)$ and $\hat{g}_i : \hat{\Omega} \to \mathbb{R}^{p+1}$, $\hat{g}_i(x) = g_i(x + x*)$. Then $D\hat{J}(0) = DJ(x_*)$, $D\hat{g}_i(0) = Dg_i(x*)$ and minimizing J in Ω with constraint $g_i = 0$ is equivalent to minimizing $\hat{J}$ in $\hat{\Omega}$ with constraint $\hat{g}_i = 0$.

$D\tilde{F}(x) = DF(x)_{|E} \in \mathcal{L}(E, \mathbb{R}^{p+1})$. 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.
As $\tilde{F} : \mathcal{U} \cap E \to \mathbb{R}^{p+1}$ is differentiable and of class C^1 since DF is continuous in $\mathcal{U}$ and for $x, y \in \mathcal{U} \cap E$,

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

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,

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

contradicts the fact that there is not open subset of $F(\mathcal{U})$ containing $F(0)$ and $DF(0)$ is not surjective. Eventually, Im $DF(0)$ is a strict subset of $\mathbb{R}^{p+1}$ and is therefore contained in some hyperplane of $\mathbb{R}^{p+1}$. There exists $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p+1}$ such that for all $h \in V$, $< \Lambda, DF(0) >= 0$, i.e.

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

As $\lambda_0 = 0$ would imply that $(Dg_1(0), \ldots, Dg_p(0))$ are linearly dependent, we conclude that $\lambda_0 \neq 0$ and divide by it. □

Exercise 9. Maximize in $\mathbb{R}^2$, $x^4 + y^4$ 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

$$E_0 = \left\{ u \in \mathrm{C}^2([a, b], \mathbb{R}) \,:\, f(a) = f(b) = 0 \right\} ,$$

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

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

- Show that J is differentiable and compute its differential.
- Show that if u_* minimizes J in E_0 then for all $t \in [a, b]$,

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

- 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 $t \in [a, b]$,

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

- 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 $t \in [a, b]$,

$$(L + \lambda K)(u(t), u'(t)) - u'(t)\partial_3 (L + \lambda K)(u(t), u'(t)) = \mu .$$

- Apply it to the Catenary's problem.
- Apply it to Dido 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 the segment $[A, B]$ is maximal.*

  Show that such a curve has constant curvature.

4.2. **Inequality constraints.** In this section V has finite dimension and we assume that

$$A = \{x \in V \,:\, g_1(x) = \ldots = g_p(x) = 0 \text{ and } h_1(x) \leq 0, \ldots, h_q(x) \leq 0\}\,,$$

with $g_1, \ldots, g_p, h_1, \ldots, h_q : V \to \mathbb{R}$ of class C^1. Let us decompose A into $A = M \cap C$ where

$$M = \{x \in V \,:\, g_1(x) = \ldots = g_p(x) = 0\} \quad \text{and} C = \{x \in V \,:\, h_1(x) \leq 0, \ldots, h_q(x) \leq 0\}\,.$$

- Assuming that at some point $x \in A$, the vectors $(\nabla g_1(x), \ldots, \nabla g_p(x))$ are linearly independdent, then M is a sub-manifold of V in a neighbourhood of x and

$$T_x M = \bigcap_{j=1}^{p} \ker Dg_j(x)\,.$$

  From the case of equality constraint, we know that the first optimality condition at a point $x_* \in A$ will write as $DJ(x_*)(h) = 0$ for all $h \in T_{x_*}M \cap ?$ where ? depends on the inequlity constraints ...
- With respect to the inequality constraints, let $x \in A$ and $i \in \{1, \ldots, q\}$, if $h_i(x) < 0$, then it will be possible to make variations in all possible directions h around x while keeping $h_i(x+h) < 0$ for h small enough. Loosely speaking, the constraint $h_i \leq 0$ is not seen from such an x, such constraints are said *inactive*, see Definition 11. If now $h_i(x) = 0$, only somme "inner" directions $h \in V$ will satisfy $h_i(x+h) \leq 0$ for h small enough. Proposition 21 states that if $< \nabla h_i(x), h >< 0$, then h is an "inner" direction.

Definition 11 (Active constraints)**.** *Let $x \in A$, the set*

$$I(x) = \{i \in \{1, \ldots, q\} \,:\, h_i(x) = 0\}$$

is called the set of active constraints *at x.*

Proposition 21 (First order condition of optimality: geometric part)**.** *Assume that $g_1, \ldots, g_p, h_1, \ldots h_q$, A, M and C are defined as above. Let $x_* \in A$ and assume that the vectors $(\nabla g_1(x_*), \ldots, \nabla g_p(x_*))$ are linearly independent. Let $J : V \to \mathbb{R}$ of class C^1. Assume that x_* is a local minimizer of J in A, then,*

$$\bigcap_{j=1}^{p} \ker Dg_j(x_*) \cap \{h \in V \,:\, \forall i \in I(x_*), < \nabla h_i(x_*), h >< 0\} \subset \{h \in V \,:< \nabla J(x_*), h >\geq 0\}\,. \tag{12}$$

Proof. As the vectors $(\nabla g_1(x_*), \ldots, \nabla g_p(x_*))$ are linearly independent, M is locally a sub-manifold around x_* and $T_{x_*}M = \bigcap_{j=1}^{p} \ker Dg_j(x_*)$.
Let $h \in T_{x_*}M$ such that for all $i \in I(x)$, $< \nabla h_i(x_*), h >< 0$, we have to prove that $< \nabla J(x_*), h >\geq 0$. Let $\gamma :]-\epsilon, \epsilon[\to M$ curve of class C^1 such that $\gamma(0) = x_*$ and $\gamma'(0) = h$. Let us show that

$$\exists \delta > 0,\, \forall t \in [0, \delta],\, \gamma(t) \in C \,(\Leftrightarrow\, \forall i \in \{1, \ldots, q\},\, h_i(\gamma(t)) \leq 0)\,.$$

Let $i \in \{1, \ldots, q\}$.
Case $i \notin I(x_*)$**:** In this case, $h_i(\gamma(0)) = h_i(x_*) < 0$ and by continuity of $h_i \circ \gamma$, $\exists \delta_i > 0$ such that for all $t \in [0, \delta_i]$, $h_i \circ \gamma(t) < 0$.
Case $i \in I(x_*)$**:** Let us write

$$\gamma(t) = \gamma(0) + t\gamma'(0) + tr(t) = x_* + t(h + r(t))\,,$$

with $r :]-\epsilon, \epsilon[\to \mathbb{R}$ such that $\lim_{t \to 0} r(t) = 0$. Then, as $h_i(x_a st) = 0$,

$$\begin{aligned} h_i(\gamma(t)) &= h_i(x_* + \underbrace{t(h + r(t))}_{O(t)}) \\ &= h_i(x_*) + < \nabla h_i(x_*), t(h + r(t)) > + o(t) \\ &= t < \nabla h_i(x_*), h > + \underbrace{t < \nabla h_i(x_*), r(t)) >}_{o(t)} + o(t) \\ &= t \underbrace{< \nabla h_i(x_*), h >}_{<0} + o(t)\,. \end{aligned}$$

Therefore, $\exists \delta_i > 0$ such that for all $t \in [0, \delta_i]$, $h_i(\gamma(t)) < 0$.

Take $\delta = \min_{i\in\{1,\dots,q\}} \delta_i$, then $\forall i \in \{1,\dots,q\}$, $\forall t \in [0,\delta]$, $h_i(\gamma(t)) \le 0$ that is $\gamma(t) \in C$, as $\gamma(t) \in M$ (by definition of γ), then $\gamma(t) \in A$. By local minimality of x_* in A and continuity of γ, up to decreasing $\delta > 0$, for all $t \in [0,\delta]$,

$$J(\gamma(t)) \ge J(*) = J(\gamma(0))\,.$$

As $J(\gamma(t)) - J(\gamma(0)) = t(J\circ\gamma)'(0) + o(t)$ and $(J\circ\gamma)'(0) =< \nabla J(\gamma(0)), \gamma'(0) >=< \nabla J(x_*), h >$, we have

$$t < \nabla J(x_*), h > +o(t) \ge 0 \quad \Rightarrow \quad \lim_{t\to 0, t>0} < \nabla J(x_*), h > \ge 0\,.$$

□

In order to pass from (**??**) to something more easy to handle, we need some "algebraic" result (as in the case of equality constraints). The result we need is known as Farkas Lemma:
...
The end of this section is "under construction" ...

Definition 12 (Qualification)**.** *The constraints are said to be* qualified *at $x \in A$ if*

$$\text{the vectors } (\nabla g_1(x),\dots,\nabla g_p(x)) \text{ are linearly independent,} \tag{13}$$

and if there exists a direction $h \in V$ such that for all i, for all $j \in I(x)$,

$$\langle \nabla g_i(x), h\rangle = 0 \text{ and } \langle \nabla h_j(x), h\rangle < 0\,. \tag{14}$$

Remark 12. A stronger assumption implying that the constraints are qualified at x is:

$$\text{the vectors } (\nabla g_i(x), \nabla h_j(x))_{i=1\dots p, j\in I(x)} \text{ are linearly independant.}$$

Theorem 22 (Karush-Kuhn-Tucker)**.** *Let V be a finite vector space and $J, g_1,\dots,g_p,h_1,\dots,h_q : V \to \mathbb{R}$. Let $x_* \in A$, assume that J, g_i, h_j are C^1 in a neighbourhood of x_* and that x_* is a relative minimum of J in A where the constraints are qualified (*(13) *and* (14)*).*
Then, there exist Lagrange multipliers $\lambda_1,\dots,\lambda_p,\mu_1,\dots,\mu_q \in \mathbb{R}$ *such that*

(1)

$$\nabla J(x_*) + \sum_{i=1}^{p} \lambda_i \nabla g_i(x_*) + \sum_{j=1}^{q} \mu_j \nabla h_j(x_*) = 0\,; \tag{15}$$

(2) $\mu_j \ge 0$ *for all* $j = 1,\dots,q$ *;*
(3) $\mu_j = 0$ *if* $h_j(x_*) < 0$ *(or equivalently here* $\mu_j h_j(x_*) = 0$*).*

Example 3. Minimize $f(x,y) = -x + y$ under the inequality constraints $y \ge x^2$ and $x + y \le 1$.

Let us define in $\mathbb{R}^2$, $h_1(x,y) = x^2 - y$ and $h_2(x,y) = x + y - 1$ and

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

Existence: f is continuous in A compact.
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. It remains to check the case where $h_1(x,y) = h_2(x,y) = 0$ whose solutions 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. We can conclude that the constraint are qualified at every point of A.
K.K.T. conditions: if (x,y) minimizes f in A, there exists $\lambda \ge 0$, $\mu \ge 0$ such that

$$\begin{cases} \nabla f(x,y) + \lambda \nabla h_1(x,y) + \mu h_2(x,y) = 0 \\ \lambda h_1(x,y) = 0 \quad \text{et} \quad \mu h_2(x,y) = 0 \\ h_1(x,y) \le 0 \quad \text{et} \quad h_2(x,y) \le 0 \end{cases} \Leftrightarrow \begin{cases} -1 + 2\lambda x + \mu = 0 \\ 1 - \lambda + \mu = 0 \\ \lambda(x^2 - y) = 0 \\ \mu(x + y - 1) = 0 \\ x^2 \le y \quad \text{et} \quad x + y \le 1 \end{cases}$$

- If $\lambda = 0$, then $1 = \mu = -1$ which is impossible.
- If $\lambda \neq 0$ and $\mu = 0$, then $\lambda = 1$ and then $x = \frac{1}{2}$ and $y = x^2 = \frac{1}{4}$. $\left(\frac{1}{2}, \frac{1}{4}\right) \in A$ 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

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

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 $2-\sqrt{5} > -\frac{1}{4}$.

Eventually, there is a unique minimizer of f in $A\ \left(\frac{1}{2}, \frac{1}{4}\right)$.

the case of convex functions ...

5. Numerical algorithms: descent methods

In this section, we study iterative algorithm to solve unconstrained and then constrained optimization problems in a finite vector space $V = \mathbb{R}^n$. Let $J : \mathbb{R}^n \to \mathbb{R}$ be smooth (at least C^1) and assume that $x_* \in \mathbb{R}^n$ be a solution of $J(x_*) = \min J$. Our purpose is to compute numerically x_*.

5.1. The 1–dimensional case.

5.1.1. *Dichotomy.* Assume that f is unimodal in $[a,b] \subset \mathbb{R}$, that is, f strictly decreasing on $[a, x_*[$ and strictly increasing on $]x_*, b]$.
Divide $[a,b]$ into 4 intervals of same length and depending on the value of f at a, b and the 3 intermediate points, find an interval of length $\frac{b-a}{2}$ containing x_* and iterate.

$$|x_{k+1} - x_*| \leq \frac{1}{2}|x_k - x_*|\,.$$

The convergence is linear.

5.1.2. *Golden section search.* Linear convergence.

5.1.3. *Newton method.* Newton's method is generally used to find a zeros of functions. The idea is to linearise

$$f(x) \simeq f(x_k) + (x - x_k) f'(x_k)\,,$$

and to define

$$x_{k+1} = x_k + \frac{f(x_k)}{f'(x_k)}\,.$$

In order to minimize f, Newton's method is applied to f' and

$$x_{k+1} = x_k + \frac{f'(x_k)}{f''(x_k)}\,.$$

When the method converges, the order is quadratic.

5.1.4. *Secant method.* As in Newton method but $f'(x_k)$ is approximated by

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\,.$$

For the minimization problem, this must be done on f' and f''.

5.2. Descent method: principle. Descent Method[General Principle].

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

5.2.1. *Relaxation method.* A natural choice of directions d_k consists in taking successively the n canonical directions $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, even though if such, d_k has no reason to be a descent direction (in the sense of Definition 13) at each step. The algorithm is thus

Algorithm 1: Relaxation method

- Choose $x_0 \in \mathbb{R}^n$.
- Given x_k, x_{k+1} is constructed 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
  $$J(x_{k+1,1} + \rho_{k,2} e_2) = \min_{\rho \in \mathbb{R}} J(x_{k+1,1} + \rho e_2) \, ;$$
  - $\ldots$;
  - $x_{k+1} = x_{k+1,n} = x_{k,n-1} + \rho_{k,n} e_n$ with $\rho_{k,n}$ such that
  $$J(x_{k+1,n-1} + \rho_{k,n} e_n) = \min_{\rho \in \mathbb{R}} J(x_{k+1,n-1} + \rho e_n) \, .$$

5.2.2. *Descent direction.*

Definition 13 (Descent direction)**.** *We say that $d \in \mathbb{R}^n$ is a (strict)* descent 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.

Proposition 23. *Let $x, d \in \mathbb{R}^n$.*

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

Proof.
- Let $h : t \in \mathbb{R} \mapsto J(x + td)$. We have that $h'(t) = \langle \nabla J(x + td), d \rangle$ so that $h'(0) = \langle \nabla J(x), d \rangle < 0$. And moreover,
$$h'(0) = \lim_{t \to 0 \ t>0} \frac{J(x + td) - J(x)}{t} \, ,$$
and consequently, there exists $t_0 > 0$ such that $\forall 0 < t < t_0$,
$$\frac{J(x + td) - J(x)}{t} < 0 \quad \Rightarrow \quad J(x + td) < J(x)$$
and thus d is a descent direction.
- As $\langle \nabla J(x), -\nabla J(x) \rangle = -\|\nabla J(x)\|^2 < 0$ ($\nabla J(x) \neq 0$) it is a descent direction by the previous point. It is the steepest descent direction meaning that $-\frac{d}{dt} J(x + td)$ is the largest possible. By Cauchy-Schwartz,
$$-\frac{d}{dt} J(x + td) = \langle \nabla J(x), d \rangle \leq \|\nabla J(x)\| \, \|d\|$$
with equality if and only if d and $\nabla J(x)$ are positively dependent.

□

Remark 13. The gradient is orthogonal to the level-sets of J.

Example 4. Let $J(x, y) = x^2 + \eta y^2$, $\eta > 1$.

There exist different descent method depending on the choice of the descent direction d_k and the step size ρ_k.

5.2.3. *Elliptic functional.* We introduce the class of *elliptic* functionals which is both well suited to study the convergence of minimization algorithm and large enough to embrace a large part of cost functions.

Definition 14 (Elliptic functional)**.** *A functional $J : V \to \mathbb{R}$ defined on a Hilbert space $(V, < \cdot, \cdot >)$ is called* elliptic *if it is C^1 and if there exists $\alpha > 0$ such that*

$$\langle \nabla J(y) - \nabla J(x), y - x \rangle \geq \alpha \|y - x\|^2 \text{ for all } x, y \in V \, .$$

Remark 14. A functional $J : V \to \mathbb{R}$ which is twice differentiable is elliptic if and only if for every $x, y \in V$,

$$\left\langle \nabla^2 J(x)y, y \right\rangle \geq \alpha \|y\|^2 .$$

Proposition 24. *Let $J : V \to \mathbb{R}$ be an elliptic functional. Then J is coercive and strictly convex.*

Corollary 25. *Let $J : V \to \mathbb{R}$ be an elliptic functional and let $A \subset V$ be a convex set. Consider the minimization problem find $x_* \in A$ such that*

$$J(x_*) = \min_{x \in A} J(x) . \tag{16}$$

Then,

- *$x_* \in A$ is solution of the problem (16) if and only if for all $x \in A$,*

$$\langle \nabla J(x_*), x - x_* \rangle \geq 0 ;$$

- *if A is closed, the problem (16) has a unique solution.*

Proof. It is an immediate consequence of Proposition 24, Theorem 10 and Theorem 17. □

Theorem 26. *If the functional $J : \mathbb{R}^n \to \mathbb{R}$ is elliptic, the relaxation method converges.*

5.3. **Gradient Methods.** As $-\nabla J(x)$ is the steepest descent direction, we choose $d_k = -\nabla J(x_k)$. It remains to choose the size step ρ_k.

5.3.1. *Gradient method with fixed step size step.*

Theorem 27 (Gradient with fixed step size)**.** *Let $J : \mathbb{R}^n \to \mathbb{R}$ be an elliptic (with ellipticity constant α) function and assume that there exists $M > 0$ such that for every $x, y \in V$,*

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

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

- *the gradient with fixed step size ρ converges to the unique solution of the minimization problem x_*;*
- *moreover, the convergence is of order one:*

$$\|x_{k+1} - x_*\| \leq \beta \|x_k - x_*\| \quad \text{with} \quad \beta = \sqrt{1 - 2\alpha\rho + M^2\rho^2} < 1 .$$

Remark 15 (Gradient with variable step). Choosing variable step sizes ρ_k such that there exist $a, b > 0$ such that $0 < a \leq \rho_k \leq b < \dfrac{2\alpha}{M^2}$, the conclusions of Theorem 27 still hold with

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

Remark 16 (Case of a quadratic functional). Let A be a n by n symmetric, positive definite matrix, $b \in \mathbb{R}^n$ and J be the quadratic functional defined in $\mathbb{R}^n$ as $J(x) = \dfrac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$. Let $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of A then J satisfies the assumptions of Theorem 27 with $\alpha = \lambda_1$ and $M = \lambda_n$ and thus $\beta = \dfrac{2\lambda_1}{\lambda_n^2}$.

Exercise 12. In the case of a quadratic functional, there is actually a better choice of step size $\rho = \dfrac{2}{\lambda_1 + \lambda_n}$ and then $\beta = \dfrac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$.

Lemma 28. *Let J be α–elliptic, then J is coercive, strictly convex and has a unique minimizer $x_* \in \mathbb{R}^n$. Moreover, for all $x, y \in \mathbb{R}^n$,*

$$J(y) - J(x) \geq \langle \nabla J(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 . \tag{17}$$

Proof. We begin with the estimate (17). Let $x, y \in \mathbb{R}^n$ and apply Taylor with integral rest formula.

$$\begin{aligned} J(y) - J(x) &= \int_{t=0}^{1} \langle \nabla J(x + t(y-x)), y - x \rangle \, dt \\ &= \langle \nabla J(x), y - x \rangle + \int_{t=0}^{1} \{ \langle \nabla J(x + t(y-x)), y - x \rangle - \langle \nabla J(x), y - x \rangle \} \, dt \\ &\geq \langle \nabla J(x), y - x \rangle + \int_{t=0}^{1} \langle \nabla J(x + t(y-x)) - \nabla J(x), t(y-x) \rangle \, \frac{dt}{t} \\ &\geq \langle \nabla J(x), y - x \rangle + \underbrace{\int_{t=0}^{1} \alpha t^2 \|y-x\|^2 \, \frac{dt}{t}}_{= \frac{\alpha}{2}\|y-x\|^2} \end{aligned}$$

In particular J is strictly convex thanks to Proposition 16 and J is coercive. Indeed,

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

Therefore J has a unique minimizer x_* in $\mathbb{R}^n$. □

Proof of Theorem 27. As $\nabla J(x_*) = 0$, then

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

Therefore,

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

□

5.3.2. *Gradient method with optimal step. Descent method with optimal step size:* A natural choice for ρ_k is choose the optimal one , that is ρ_k solution to the 1–dimensional minimization problem

$$J(x_k + \rho_k d_k) = \min_{\rho \in \mathbb{R}} J(x_k + \rho d_k) \, .$$

This method reduces a n–dimensional optimisation problem to successive 1–dimensional problems. *Gradient with optimal step size:* $d_k = -\nabla J(x_k)$.

Theorem 29 (Gradient with optimal step size)**.** *If J is elliptic, then the gradient with optimal step size method is well-defined at each step ($\forall k$, $\exists \rho_k$) and converges to x_*.*

Remark 17. In the case where $J(x) = \frac{1}{2}a(x,x) - b(x)$ is a quadratic elliptic functional, once d_k is fixed, determining ρ_k such that

$$\begin{aligned} (18) \quad J(x_k + \rho_k d_k) &= \min_{\rho \in \mathbb{R}} J(x_k + \rho d_k) \\ &= \min_{\rho \in \mathbb{R}} \left(\frac{1}{2} a(d_k, d_k) \rho_k^2 + (a(x_k, d_k) - b(d_k)) \, \rho_k + \frac{1}{2} a(x_k, x_k) - b(x_k) \right) \end{aligned}$$

is direct.

Remark 18 (Orthogonality of successive directions). Let $h_k(\rho) = J(x_k + \rho d_k) = J(x_k - \rho \nabla J(x_k))$, h_k is C^1 and

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

Consequently, two successive descent directions are orthogonal.

The main drawback of this methods is that there is a 1–dimensional minimization problem to solve at each step. Moreover, it is pointless to optimize the step size if the descent direction is "bad", indeed the choice $d_k = -\nabla J(x_k)$ relies on a local estimate.

5.3.3. *Projected gradient: handling convex constraints.* From a practical point of view, it essentially allows to handle constraints of the form $A = \prod_{i=1}^n [a_i, b_i]$. The principle is the same as for any descent method, but at each step, there is an additional projection onto the constraints set. Let A be the closed convex set of constraints and P be the associated projection onto A. Then x_{k+1} is defined as

$$x_{k+1} = P(x_k + \rho_k d_k) \, .$$

The problem is that this projector is generally difficult to compute. There is at least one case where P has a simple expression: for $V = \prod_{i=1}^n [a_i, b_i]$ then

$$(P(x))_i = \min\{\max(a_i, x_i), b_i\} = \left\{ \begin{array}{lll} a_i & \text{if} & x_i < a_i \\ x_i & \text{if} & x_i \in [a_i, b_i] \\ b_i & \text{if} & x_i > b_i \end{array} \right. .$$

For the projected gradient with variable step, the same convergence property holds as for the gradient with fixed/variable step.

Theorem 30 (Projected gradient with fixed step size)**.** *Let A be a closed convex set and $J : A \subset \mathbb{R}^n \to \mathbb{R}$ be an elliptic (with ellipticity constant α) function and assume that there exists $M > 0$ such that for every $x, y \in V$,*

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

Then, if there exist $a, b > 0$ such that for all k, $0 < a \leq \rho_k \leq b < \dfrac{2\alpha}{M^2}$,

- *the projected gradient with variable step size ρ_k converges to the unique solution of the minimization problem x_*;*
- *moreover, the convergence is of order one:*

$$\|x_{k+1} - x_*\| \leq \beta \|x_k - x_*\| \quad \textit{with} \quad \beta = \sqrt{1 - 2\alpha\rho + M^2\rho^2} < 1 \, .$$

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

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

which characterize x_* as the projection on A of $x_* - \rho_k \nabla J(x_*)$, that is $P(x_* - \rho_k \nabla J(x_*)) = x_*$. Consequently, following the proof in the case without constraints and using the fact that P is 1–lipschitz,

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

□

Example 5. Let $A = \overline{B(0,1}$ closed unit ball of $\mathbb{R}^n$, then

$$P(x) = \left\{ \begin{array}{ll} x & \text{if} \quad \|x\| \leq 1 \\ \frac{x}{\|x\|} & \text{otherwise} \end{array} \right.$$

5.3.4. *Newton's method.* **Idea:** use Newton's method in order to solve $x_* \in \mathbb{R}^n$ such that $\nabla J(x_*) = 0$.

$$x_{k+1} = x_k - \left(\nabla^2 J(x_k)\right)^{-1} \nabla J(x_k) \, .$$

Generalized Newton methods:

$$x_{k+1} = x_k - \left(A(x_k)\right)^{-1} \nabla J(x_k) \, .$$

with $A(x_k) = \rho_k Id$, we recover gradient methods.

5.3.5. *Penalization method.*

5.4. **The case J quadratic.**

5.4.1. *Failure of gradient methods.*

5.4.2. *Conjugated gradient.*

6. Duality and Uzawa algorithm

In this section, V, W are finite vector spaces, even though it is possible to state more general results in Hilbert spaces.

6.1. **Introduction to duality.** Let us come back to the problem of minimization under inequality constraints:

$$J(x_*) = \min_{x\in A} J(x) \tag{19}$$

with

$$A = \{x \in V \,:\, h_1(x) \leq 0, \dots, h_q(x) \leq 0\} \ .$$

Let us denote $h : V \to \mathbb{R}^q$, $x \mapsto (h_1(x), \dots, h_q(x))$, for the sake of simplicity, J and h are assumed to be defined and C^1 in V. If we introduce the Lagrangian

$$\begin{array}{rccl} L & : & V \times (\mathbb{R}_+)^q & \to & \mathbb{R} \\ & & (x,\mu) & \mapsto & J(x) + \sum_{i=1}^q \mu_i h_i(x) \ , \end{array}$$

then, if x_* is a minimizer (and the qualification condition satisfied at x_*), the conclusion of Kuhn–Tucker Theorem implies that there exists $\mu_* \in (\mathbb{R}_+)^q$ such that (x_*, μ_*) satisfy

$$\begin{aligned} \nabla_x L(x_*,\mu_*) &= \nabla J(x_*) + \sum_{i=1}^q \mu_{*,i} \nabla h_i(x_*) = 0 \\ \nabla_\mu L(x_*,\mu_*) &= h(x_*) = 0 \ . \end{aligned}$$

And thus (x_*, μ_*) is a critical point of L. Actually, under some additional convexity assumptions, it is possible to be more precise on the type of this critical point.

Theorem 31.
- *If $(x_*, \mu_*) \in V \times (\mathbb{R}_+)^q$ is a saddle point of L then $x_* \in A$ and is a solution to the constrained minimization problem* (19).
- *Conversely, if h_i are convex in A and the constraints are qualified in A, then if $x_a st$ is a solution to the minimization problem* (19)*, there exists $\mu_* \in (\mathbb{R}_+)^q$ such that (x_*, μ_*) is a saddle point of L.*

Definition 15 (Saddle point)**.** *Let $X \subset V$ and $M \subset W$ and $L : X \times M \to \mathbb{R}$. A point $(x, \mu) \in X \times M$ is said to be a* saddle point *of L in $X \times M$ if*

- *x minimizes $L(\cdot, \mu)$ in X and,*
- *μ minimizes $L(x, \cdot)$ in M.*

In other words,

$$\sup_{\mu\in Y} L(x_*,\mu) = L(x_*,\mu_*) = inf_{x\in X} L(x,\mu_*) \ .$$

Theorem 32 (Duality)**.** *(x_*, μ_*) is a saddle point of L in $X \times M$ if and only if*

$$\sup_{\mu\in M} \inf_{x\in X} L(x,\mu) = L(x_*,\mu_*) = \inf_{x\in X} \sup_{\mu\in M} L(x,\mu) \ .$$

Remark 19. Notice that

$$\sup_{\mu\in M} \inf_{x\in X} L(x,\mu) \leq \inf_{x\in X} \sup_{\mu\in M} L(x,\mu)$$

is always true.

If we denote

$$\begin{aligned} \mathcal{I}(x) = \sup_{\mu\in M} L(x,\mu) \text{ and } \mathcal{G}(x) = \inf_{x\in X} L(x,\mu) \ , \\ \mathcal{I}(x_*) = \min_{x\in V} \mathcal{I}(x) \text{ is called the } \textit{primal} \text{ problem} \end{aligned}$$

and

$$\mathcal{G}(\mu_*) = \max_{\mu\in M} \mathcal{G}(\mu) \text{ is called the } \textit{dual} \text{ problem} \ .$$

In the case where $L(x,\mu) = J(x) + \sum_{i=1}^q \mu_i h(x)$ is defined in $V \times (\mathbb{R}_+)^q$, and under the assumptions of Theorem 31, we know that x_* minimizes J in A if and only if there exists $\mu_* \in (\mathbb{R}_+)^q$ such that (x_*, μ_*) is a saddle point of L in $V \times (\mathbb{R}_+)^q$. Therefore, (x_*, μ_*) is solution of the primal problem if and only if it is solution of the second problem.
Notice that the primal problem is exactly the inequality constrained problem as

$$\sup_{\mu\in(\mathbb{R}_+)^q} J(x) + \sum_{i=1}^q \mu_i h(x) = \begin{cases} J(x) & \text{if} \quad x \in A \\ +\infty & \text{otherwise.} \end{cases}$$

While the dual problem is generally easier to handle since there is no constraint in $x \in V$ and only the simple constraints $\mu_i \geq 0$. This is the idea of Uzawa algorithm: solve the dual problem rather than the primal.

6.2. **Uzawa algorithm.** Uzawa algorithm is just the projected gradient method applied to the dual problem. The projection operator $P_+ : \mathbb{R}^q \to (\mathbb{R}_+)^q$ is simply computed as

$$P_+(\mu = (\mu_1, \ldots, \mu_q)) = (\max(\mu_i, 0))_i \ .$$

Algorithm 2: Uzawa algorithm

- Start from $\mu_0 \in (\mathbb{R}_+)^q$.
- Given x_{k-1} and μ_k, x_k and μ_{k+1} are constructed as follows.
  - x_k is a minimizer of the unconstrained problem $\min_{x \in V} L(x, \mu_k) = \mathcal{G}(\mu_k)$;
  - Choose ρ_k;
  - Set $\mu_{k+1} = P_+(\mu_k + \rho_k \nabla \mathcal{G}(\mu_k))$ where

$$\nabla \mathcal{G}(\mu_k) = (h_i(x_k))_i \ .$$

Remark 20. The fact that x_k indeed tends to a solution of the primal problem remains to check and the formula giving the gradient of $\mathcal{G}$ holds only under suitable assumptions, as the continuous dependence of $x_\mu = \min_{x \in V} L(x, \mu)$ with respect to μ.

7. Corrections

Correction 5. Let $A \subset V$ be a convex set. Check that $J : A \to \mathbb{R}$ is convex $\Leftrightarrow$ epi(J) is convex.

It is a simple application of definitions. Assume first that J is convex. Let $(x_1, \lambda_1), (x_2, \lambda_2) \in \text{epi}(J)$ and $t \in [0,1]$, then $t(x_1, \lambda_1) + (1-t)(x_2, \lambda_2) \in \text{epi}(J)$. Indeed, for $i = 1, 2$, $\lambda_i \geq J(x_i)$ so that, using the convexity of J,

$$t\lambda_1 + (1-t)\lambda_2 \geq tJ(x_1) + (1-t)J(x_2) \geq J(tx_1 + (1-t)x_2) .$$

The converse property has a similar proof. Assume now that epi(J) is convex. Let $x_1, x_2 \in A$, $t \in [0,1]$. As $(x_1, J(x_1)), (x_2, J(x_2)) \in \text{epi}(J)$ and epi(J) is convex, then $(tx_1+(1-t)x_2, tJ(x_1)+(1-t)J(x_2)) \in \text{epi}(J)$. Consequently, $tJ(x_1) + (1-t)J(x_2) \geq J(tx_1 + (1-t)x_2, tJ(x_1) + (1-t)J(x_2))$. And we proved that J is convex.

Correction 6. We consider the Hilbert (thus reflexive) space

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

provided with the scalar product $(x_n)_n \cdot (y_n)_n = \sum_{n=0}^{\infty} x_n y_n$, and we define

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

Check that f is coercive and lower semi continuous and check that however, f does not admit minimizer on $l^2(\mathbb{R})$.

- f is coercive:

$$f(x) \geq \left(\|x\|^2 - 1\right)^2 \xrightarrow[\|x\|\to+\infty]{} +\infty .$$

- f is strongly l.s.c. (continuous actually): $x \mapsto \left(\|x\|^2 - 1\right)^2$ is strongly continuous and

$$\sum_{n=0}^{\infty} \frac{(x_n)^2}{n+1} = (x_n)_n \cdot L((x_n)_n)$$

  where $L((x_n)_n) = \left(\frac{x_n}{n+1}\right)_n$. As the scalar product is strongly continuous, it remains to prove that the application $L : l^2(\mathbb{R}) \to l^2(\mathbb{R})$ is continuous. As $\frac{1}{n+1} \leq 1$, for $x = (x_n)_n, y = (y_n)_n \in l^2(\mathbb{R})$ we have

$$\|L(x) - L(y)\| = \sum_{n=0}^{\infty} \frac{(x_n - y_n)^2}{(n+1)^2} \leq \sum_{n=0}^{\infty} (x_n - y_n)^2 = \|x - y\|^2 \xrightarrow[\|x-y\|\to 0]{} 0.$$

- f does not admit a minimizer: for all $x \in l^2(\mathbb{R})$, $f(x) > 0$. Indeed, let $x \in l^2(\mathbb{R})$, $f(x) \geq 0$ and if $f(x) = 0$ then $\left(\|x\|^2 - 1\right)^2 = 0$ which implies $\|x\| = 1$ and $\sum_{n=0}^{\infty} \frac{(x_n)^2}{n+1}$ which implies $x = 0$. Let us now show that $\inf_{l^2(\mathbb{R})} f = 0$. For $k \in \mathbb{N}$, define the sequence $x^k \in l^2(\mathbb{R})$ by $x_n^k = \delta_{n,k}$ so that $\|x^k\| = 1$ and $f(x^k) = \frac{1}{k+1} \xrightarrow{k\to\infty} 0$.

We can deduce from the previous properties that f is not weakly l.s.c. in $l^2(\mathbb{R})$.

Correction 7. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and consider

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

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

Let $a, b \in \mathbb{R}$ and $t \in [0,1]$ and consider a 1–periodic function $\psi : \mathbb{R} \to \mathbb{R}$ s. t. $\psi(x) = \begin{array}{ccc} a & \text{if} & 0 \leq x < t \\ b & \text{if} & t \leq x \leq 1 \end{array}$. Let then $u_k(x) = \psi(kx)$ for $k \in \mathbb{N}^*$.

- For all $k \in \mathbb{N}^*$, $\displaystyle\int_0^1 f(u_k) = tf(a) + (1-t)f(b)$. Indeed, thanks to the change of variable $y = kx$ and periodicity of ψ, we get

$$\int_0^1 f(u_k) = \int_{x=0}^1 f(\psi(kx))\,dx = \frac{1}{k}\int_{y=0}^k f(\psi(y))\,dy = \frac{1}{k}k\int_0^1 f(\psi) = tf(a) + (1-t)f(b)\,.$$

- The sequence $(u_k)_k$ $\mathrm{L}^2([0,1])$–weakly converges to the constant function $ta + (1-t)b$. Indeed, from the previous point with $f(z) = z^2$, we have that $\|u_k\|^2_{\mathrm{L}^2} = ta^2 + (1-t)b^2$. therefore $(u_k)_k$ is uniformly bounded in $\mathrm{L}^2([0,1])$ and thus there is a subsequence L^2–weakly converging to $u \in \mathrm{L}^2([0,1])$. Let $v \in \mathrm{C}([0,1])$,

$$\begin{aligned}\int_0^1 u_k v = \frac{1}{k}\int_{y=0}^k \psi(y)v(y/k)\,dy = \frac{1}{k}a\sum_{i=0}^{k-1}\int_i^{i+t} v(y/k)dy + \frac{1}{k}b\sum_{i=0}^{k-1}\int_{i+t}^{i+1} v(y/k)dy \\ = a\sum_{i=0}^{k-1}\int_{\frac{i}{k}}^{\frac{i+t}{k}} v(y)dy + b\sum_{i=0}^{k-1}\int_{\frac{i+t}{k}}^{\frac{i+1}{k}} v(y)dy\,.\end{aligned}$$

  For every $k \in \mathbb{N}^*$, there exists (mean value theorem) $c_k \in \left[\frac{i}{k}, \frac{i+t}{k}\right]$ such that

$$\frac{k}{t}\int_{\frac{i}{k}}^{\frac{i+t}{k}} v(y)dy = v(c_k)\,.$$

  By generalized Riemann sums theorem, we have $\displaystyle\frac{1}{k}\sum_{i=0}^{k-1} v(c_k) \to \int_0^1 v$ and therefore

$$\int_0^1 u_k v \to \int_0^1 (ta + (1-t)b)v.$$

  We can conclude that $u = ta + (1-t)b$. In order to obtain the weak convergence of the whole sequence, it is enough to use the density of $\mathrm{C}([0,1])$ in $\mathrm{L}^2([0,1])$ (for the L^2–norm) and the fact that $\|u_k - u\|_{\mathrm{L}^2} \le \|u\|_{\mathrm{L}^2} + ta^2 + (1-t)b^2$ is bounded independently of k.[7]

If now L is L^2–weakly l.s.c., then

$$f(ta + (1-t)b) = \int_0^1 f(u) \le \liminf_{k\infty} \int_0^1 f(u_k) = tf(a) + (1-t)f(b).$$

And f is convex.

Correction 8 (An example in infinite dimension ...). Let us consider the vector space

$$E_0 = \left\{u \in \mathrm{C}^2([a,b],\mathbb{R}) \,:\, f(a) = f(b) = 0\right\}\,,$$

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

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

(1) Show that J is differentiable and compute its differential.
(2) Show that if u_* minimizes J in E_0 then for all $t \in [a,b]$,

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

[7] Let $\epsilon > 0$, $v \in \mathrm{L}^2([0,1]]$ and set $C = \|u\|_{\mathrm{L}^2} + ta^2 + (1-t)b^2$. By density, let $v_\epsilon \in \mathrm{C}([0,1])$ be such that $\|v - v_\epsilon\|_{\mathrm{L}^2} \le \frac{\epsilon}{2C}$. We have

$$\left|\int_0^1 (u-u_k)v\right| \le \left|\int_0^1 (u-u_k)v_\epsilon\right| + \left|\int_0^1 (u-u_k)(v-v_\epsilon)\right| \le \underbrace{\left|\int_0^1 (u-u_k)v_\epsilon\right|}_{\le \frac{\epsilon}{2} \text{ for } k \text{ large enough}} + \underbrace{\|u-u_k\|_{\mathrm{L}^2}\|v-v_\epsilon\|_{\mathrm{L}^2}}_{\le \frac{\epsilon}{2}}$$

(3) Show that if u_* minimizes J in E_0 under the constraint $G(u) = \alpha$ with $DG(u_*) \neq 0$, then there exists $\lambda \in \mathbb{R}$ such that for all $t \in [a, b]$,

$$\frac{d}{dt}\left(\partial_3(L + \lambda K)(t, u(t), u'(t))\right) = \partial_2(L + \lambda K)(t, u(t), u'(t))\,. \tag{20}$$

(4) 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 $t \in [a, b]$,

$$(L + \lambda K)(u(t), u'(t)) - u'(t)\partial_2(L + \lambda K)(u(t), u'(t)) = \mu\,. \tag{21}$$

(5) Apply it to the Catenary's problem.

(6) Apply it to Dido 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 the segment $[A, B]$ is maximal.*

Show that such a curve has constant curvature.

(1) J and G are differentiable, indeed, from exercise 6, for all $h \in \mathrm{C}^1([a, b], \mathbb{R})$, thus for all $h \in E_0 = \mathrm{C}^2([a, b], \mathbb{R})$,

$$J(u + h) - J(u) = l_u(h) + r_u(h)\,,$$

with l_u linear and continuous, and $r_u(h) = o(\|h\|_{\mathrm{C}^1})$. As $\|u\|_{\mathrm{C}^1} \leq \|u\|_{\mathrm{C}^2}$, $l_u : \mathrm{C}^2([a, b], \mathbb{R}) \to \mathbb{R}$ is linear continuous with respect to $\|\cdot\|_{\mathrm{C}^2}$ and $r_u(h) = o(\|h\|_{\mathrm{C}^2})$. Consequently J is differentiable and $DJ(u) = l_u$ (l_u restricted to $\mathrm{C}^2([a, b], \mathbb{R})$).

In addition, J is C^1. Indeed, let $u, h \in E_0$ s.t. $\|h\|_{\mathrm{C}^2} \leq 1$, then

$$\|DJ(u + h) - DJ(u)\|_{op} \leq \int_a^b \left\|\nabla L(t, u(t) + h(t), u'(t) + h'(t)) - (t, u(t), u'(t))\right\|\, dt$$

and for all $t \in [a, b]$, $(t, u(t) + h(t), u'(t) + h'(t))$ and $(t, u(t), u'(t))$ are contained in some fixed compact set, in which ∇L is κ–Lipschitz (since C^2). Hence

$$\|DJ(u + h) - DJ(u)\|_{op} \leq \kappa \int_a^b \left\|(0, h(t), h'(t)\right\|\, dt \leq \kappa(b - a)\|h\|_{\mathrm{C}^2}\,,$$

and the continuity of DJ follows.

(2) See next question.

(3) Let us apply Lagrange multipliers theorem. J and G are C^1 and the constraints are qualified if and only if $DG(u) \neq 0$, hence there exists $\lambda \neq 0$ such that $DJ(u_*) + \lambda DG(u_*) = 0$, that is, for every $h \in E_0$,

$$\int_a^b \partial_2(L + \lambda K)(t, u(t), u'(t))h(t) + \partial_3(L + \lambda K)(t, u(t), u'(t))h'(t)\, dt = 0\,.$$

As $h(a) = h(b) = 0$, it follows from by parts integration that

$$\int_a^b \left\{\partial_2(L + \lambda K)(t, u(t), u'(t))h(t) - \frac{d}{dt}\partial_3(L + \lambda K)(t, u(t), u'(t))\right\} h(t)\, dt = 0\,.$$

Then, by the fundamental lemma of calculus of variations, (20) holds.

(4) Let us differentiate the left hand side of (21). For all $t \in [a, b]$,

$$\begin{aligned} & \frac{d}{dt}\left\{(L + \lambda K)(u(t), u'(t)) - u'(t)\partial_2(L + \lambda K)(u(t), u'(t))\right\} \\ = \;& u'(t)\partial_1(L + \lambda K)(u(t), u'(t)) - u'(t)\frac{d}{dt}\partial_2(L + \lambda K)(u(t), u'(t)) \\ = \;& 0 \end{aligned}$$

thanks to the previous question.

(5) Assume that the chain is described as the graph of a regular function u over the segment $[a, b]$. Then the potential energy of the chain is proportional to the integral of the height of the chain along the chain itself, that is

$$J(u) = \int_a^b u(t)\sqrt{1 + u'(t)^2}\, dt\,,$$

and the length constraint is then $G(u) := \int_a^b \sqrt{1+u'(t)^2}\,dt - l = 0$. The problem of minimizing J under the constraint $G(u) = 0$ falls in the scope of the previous questions when defining $L, K : \mathbb{R}^3 \to \mathbb{R}$ of class C^2 as

$$L(x_1,x_2) = x_1\sqrt{1+(x_2)^2} \quad \text{and} \quad K(x_1,x_2) = \sqrt{1+(x_2)^2} - \frac{l}{b-a}\,.$$

Therefore, if u is a minimizer of J under the constraint $G(u) = 0$, either $DG(u) = 0$ (and this corresponds to the minimal length possible $l = b-a$ and u is equally constant to 0[8]), either there exists λ, $\mu \in \mathbb{R}$ such that (21) holds. As

$$(L+\lambda K)(x_1,x_2) = (x_1+\lambda)\sqrt{1+x_2^2} - \frac{\lambda l}{b-a} \quad \text{and} \quad x_2\partial_2(L+\lambda K)(x_1,x_2) = (x_1+\lambda)\frac{x_2^2}{\sqrt{1+x_2^2}}\,,$$

and $\frac{\lambda l}{b-a}$ is a constant, (21) rewrites as $\exists \mu \in \mathbb{R}$ s.t.

$$(u+\lambda)\sqrt{1+(u')^2} - (u+\lambda)\frac{(u')^2}{\sqrt{1+(u')^2}} = \mu \quad \Leftrightarrow \quad u+\lambda = \mu\sqrt{1+(u')^2}\,.$$

Setting $v = u+\lambda$, we obtain the differential equation $v = \mu\sqrt{1+(v')^2}$. Either $\mu = 0$ and u is constant (equal to $\lambda = 0$) and $l = b-a$, or $v^2 \geq \mu^2 > 0$ and for all $t \in [a,b]$,

$$|v'(t)| = \sqrt{\left(\frac{v(t)}{\mu}\right)^2 - 1}\,.$$

On any interval where v' is non zero, we conclude that v is of the form

$$v(t) = \mu\cosh\left(\frac{t+c_1}{\mu}\right) \quad \text{or} \quad v(t) = \mu\cosh\left(\frac{-t+c_2}{\mu}\right) = \mu\cosh\left(\frac{t-c_2}{\mu}\right)\,.$$

If $v'(t) = 0$ on some interval, then $v(t) = \mu$ or $v(t) = -\mu$. When trying to glue those three possible forms of solution at some point, then continuity and C^2 regularity implie that

$$v(t) = \mu\cosh\left(\frac{t+c}{\mu}\right), \quad c \in \mathbb{R}\,.$$

As $v(a) = v(b) = \lambda$ and $a < b$, then

$$\mu\cosh\left(\frac{a+c}{\mu}\right) = \mu\cosh\left(\frac{a+c}{\mu}\right) = \lambda \quad \Rightarrow \quad a+c = b+c \text{ or } a+c = -b-c$$

whence $c = -\dfrac{a+b}{2}$ and $\lambda = \mu\cosh\left(\dfrac{b-a}{2\mu}\right)$. It remains to use the length constraint to determine μ. As $v = u+\lambda = \mu\sqrt{1+(u')^2}$,

$$l = \frac{1}{\mu}\int_a^b v(t)\,dt = \int_a^b \cosh\left(\frac{t-\frac{a+b}{2}}{\mu}\right)\,dt = \mu\left(\sinh\left(\frac{b-a}{2\mu}\right) - \sinh\left(\frac{a-b}{2\mu}\right)\right)$$
$$= 2\mu\sinh\left(\frac{b-a}{2\mu}\right)\,.$$

Eventually

$$u(t) = \mu\left(\cosh\left(\frac{t-\frac{a+b}{2}}{\mu}\right) - \cosh\left(\frac{b-a}{2\mu}\right)\right) \quad \text{with} \quad 2\mu\sinh\left(\frac{b-a}{2\mu}\right) = l\,.$$

[8]Indeed, as $\partial_1 K(u(t),u'(t)) = 0$, $DG(u) = 0$ is equivalent to $\dfrac{d}{dt}\partial_2 K(u(t),u'(t)) = 0$ so that their exists some $\beta \in \mathbb{R}$ such that

$$\frac{u'}{\sqrt{1+(u')^2}} = \beta.$$

After easy computations, this leads to $(u')^2$ constant and by continuity, u' constant. Hence u is affine and $u(a) = u(b)$ eventually imply that u is constant and $l = b-a$.