Chapter 3

Abstract variational principles

The introduction in Chapter 2 of a weak formulation of the model examples (Dirichlet problem, Stokes system) leads to study of the following problem.

Given a linear vector space V, a bilinear symmetrical form $a : V \times V \longrightarrow \mathbf{R}$, and a linear form $L : V \longrightarrow \mathbf{R}$, find $u \in V$ such that

$$a(u,v) = L(v) \quad \forall\, v \in V. \tag{3.1}$$

When a is positive, this turns out to be equivalent to the following minimization problem: find $u \in V$ such that

$$J(u) \leq J(v), \tag{3.2}$$

where $J(v) := \frac{1}{2}a(v,v) - L(v)$.

In this chapter, we introduce the topological and geometrical concepts which allow us to solve this kind of problem and much more.

3.1 ▪ The Lax–Milgram theorem and the Galerkin method

3.1.1 ▪ The Lax–Milgram theorem

In this section, V is a Hilbert space equipped with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm:

$$\forall v \in V \quad ||v||^2 = \langle v, v \rangle.$$

Let us recall the celebrated Riesz theorem.

Theorem 3.1.1 (Riesz). *Let V be a Hilbert space and $L \in V^*$ a linear continuous form on V. Then there exists a unique $f \in V$ such that*

$$\forall\, v \in V \quad L(v) = \langle f, v \rangle.$$

Notice that given $f \in V$, the linear form L_f defined by

$$L_f(v) = \langle f, v \rangle$$

satisfies (by application of the Cauchy–Schwarz inequality)

$$|L_f(v)| \leq ||f||\,||v||$$

and hence

$$||L_f||_* \leq ||f||,$$

where $||L_f||_*$ is the dual norm of the continuous linear form L_f. On the other hand, by taking $v = \frac{1}{||f||} f$ (if $f \neq 0$) we obtain

$$||L_f||_* \geq ||f||.$$

The Riesz theorem tells us that the linear isometrical embedding $f \mapsto L_f$ from V into V^* is onto. So V and V^* can be identified both as vector spaces and as Hilbert spaces.

Note that it is not completely correct to say that the topological dual of V is V itself! The Riesz theorem tells us that the topological dual of V, that is, V^*, is isometric to V and describes how any element of V^* can be uniquely represented with the help of an element of V: the mapping

$$f \in V \mapsto L_f \in V^*$$

is an isometrical isomorphism from V onto V^*.

So we will often identify V^* with V. But one may imagine other representations of V^*. We will illustrate this when describing the dual of the Sobolev space $H_0^1(\Omega)$.

An important situation where one has to be careful with such identifications is when we have an embedding (i linear continuous, V dense in H)

$$V \overset{i}{\hookrightarrow} H$$

of two Hilbert spaces, $(V, (\cdot,\cdot))$ and $(H, \langle\cdot,\cdot\rangle)$. Clearly, any linear continuous form L on H when "restricted" to V (indeed, $L_{|V} = L \circ i$) defines a linear continuous form on V. So $H^* \subset V^*$ and the mapping $L \in H^* \mapsto L_{|V} \in V^*$ is one to one because V is dense in H.

When identifying H and H^* we have the usual "triplet"

$$V \hookrightarrow H \hookrightarrow V^*.$$

But now we cannot also identify V and V^* because we should end with the conclusion that $H \hookrightarrow V$!

Thus one has to choose one identification. One cannot identify both (H and H^*) and (V and V^*). A typical situation is

$$V = H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega) = V^*.$$

This leads to a representation of the dual of the Hilbert space $V = H_0^1(\Omega)$ which is different from the Riesz representation. The Riesz representation theorem will play a key role when establishing the following theorem.

Theorem 3.1.2 (Lax–Milgram). *Let V be a Hilbert space with the scalar product $\langle\cdot,\cdot\rangle$ and $||\cdot|| = \sqrt{\langle\cdot,\cdot\rangle}$ the associated norm.*

Let $a : V \times V \longrightarrow \mathbf{R}$ be a bilinear form which satisfies (i) *and* (ii)*:*

(i) *a is continuous, that is, there exists a constant $M \in \mathbf{R}^+$ such that*

$$\forall u, v \in V \quad |a(u,v)| \leq M||u|| \cdot ||v||;$$

(ii) *a is coercive, that is, there exists a constant $\alpha > 0$ such that*

$$\forall v \in V \quad a(v,v) \geq \alpha ||v||^2.$$

Then for any $L \in V^*$ (L is a linear continuous form on V) there exists a unique $u \in V$ such that

$$a(u,v) = L(v) \quad \forall v \in V.$$

Remark 3.1.1. Let us make some complements to the discussions above.

(a) Before proving the Lax–Milgram theorem, let us notice that it contains as a particular case the Riesz representation theorem. Take $a(u,v) = \langle u,v \rangle$ and verify (i) and (ii). By the Cauchy–Schwarz inequality we have

$$|a(u,v)| \leq ||u|| \, ||v||.$$

Hence a is continuous (take $M = 1$). Moreover, $a(v,v) = \langle v,v \rangle = ||v||^2$ and a is coercive (take $\alpha = 1$). So, for any $L \in V^*$ there exists a unique $u \in V$ such that

$$L(v) = \langle u,v \rangle \quad \forall\, v \in V,$$

and this is the Riesz representation theorem.

(b) One can easily verify that for a bilinear form, the continuity property is equivalent to the existence of some constant $M \geq 0$ such that

$$|a(u,v)| \leq M ||u|| \, ||v||. \tag{3.3}$$

Let us first verify that if (3.3) is satisfied, then a is continuous: take $u_n \longrightarrow u$ and $v_n \longrightarrow v$. Then

$$\begin{aligned} a(u_n,v_n) - a(u,v) &= a(u_n,v_n) - a(u_n,v) + a(u_n,v) - a(u,v) \\ &= a(u_n, v_n - v) + a(u_n - u, v). \end{aligned}$$

It follows that

$$|a(u_n,v_n) - a(u,v)| \leq M ||u_n|| \cdot ||v_n - v|| + M ||u_n - u|| \cdot ||v||. \tag{3.4}$$

The sequence $(u_n)_{n \in \mathbf{N}}$ being norm convergent is bounded in V and there exists a constant $C \geq 0$ such that $\sup_n ||u_n|| \leq C$. Returning to (3.4),

$$|a(u_n,v_n) - a(u,v)| \leq M\big[C ||v_n - v|| + ||v|| \cdot ||u_n - u||\big],$$

which implies that $\lim_n a(u_n,v_n) = a(u,v)$.

Conversely, let us assume that a is a bilinear continuous form on $V \times V$. Since $a(0,0) = 0$ and a is continuous at $(0,0)$, for any $\varepsilon > 0$ there exists some $\eta(\varepsilon) > 0$ such that

$$||u|| \leq \eta(\varepsilon) \text{ and } ||v|| \leq \eta(\varepsilon) \Longrightarrow \; a(u,v)| \leq \varepsilon.$$

Take u, v arbitrary elements of V, $u \neq 0$, $v \neq 0$. Then

$$\left\| \frac{\eta(\varepsilon)}{||u||} u \right\| \leq \eta(\varepsilon) \quad \text{and} \quad \left\| \frac{\eta(\varepsilon)}{||v||} v \right\| \leq \eta(\varepsilon).$$

Hence

$$\left| a\left(\frac{\eta(\varepsilon)}{||u||} u, \frac{\eta(\varepsilon)}{||v||} v \right) \right| \leq \varepsilon,$$

which implies

$$\forall\, u,v \neq 0 \quad |a(u,v)| \leq \frac{\varepsilon}{\eta(\varepsilon)^2}\, ||u|| \cdot ||v||.$$

This is still true if u or v is the zero element of V. So, one can take $M = \varepsilon/\eta^2(\varepsilon)$.

PROOF OF THE LAX–MILGRAM THEOREM. When establishing a weak formulation for some partial differential equations or systems

$$Au = f \tag{3.5}$$

(for example, $Au = -\Delta u$ for the Dirichlet problem with prescribed boundary data contained in the domain of A), we have been led to study problems of the form

$$a(u,v) = L(v) \quad \forall\, v \in V. \tag{3.6}$$

Indeed, we are going to reconstruct an abstract equation (3.5) from (3.6). The major interest of this reverse operation is that now we are able to formulate precisely the topological and geometrical properties of the operator A.

Let us first apply the Riesz theorem to L: there exists some $f \in V$ such that

$$L(v) = \langle f, v\rangle \quad \forall v \in V. \tag{3.7}$$

For any fixed $u \in V$, the mapping $v \mapsto a(u,v)$ is a continuous linear form on V; note that

$$|a(u,v)| \leq M||u||\,||v|| \quad \forall v \in V.$$

Applying once more the Riesz theorem, there exists a unique element, which we denote $A(u) \in V$, such that

$$\forall v \in V \quad a(u,v) = \langle A(u), v\rangle.$$

The mapping $u \mapsto A(u)$ from V into V is linear: given u_1, u_2 belonging to V

$$\begin{aligned}\langle A(u_1+u_2), v\rangle &= a(u_1+u_2, v)\\ &= a(u_1,v) + a(u_2,v)\\ &= \langle A(u_1), v\rangle + \langle A(u_2), v\rangle\\ &= \langle A(u_1) + A(u_2), v\rangle \qquad \forall\, v \in V.\end{aligned}$$

Hence,

$$A(u_1+u_2) = A(u_1) + A(u_2).$$

Similarly,

$$\forall\, \lambda \in \mathbf{R},\ \forall\, u \in V \quad A(\lambda u) = \lambda A(u).$$

So, our problem can be reformulated as follows: find $u \in V$ such that

$$\langle A(u), v\rangle = \langle f, v\rangle \quad \forall v \in V,$$

that is,

$$Au = f. \tag{3.8}$$

Let us reformulate in terms of A the properties of the bilinear form $a(\cdot,\cdot)$:

(a) since a is bilinear we have that

$$A : V \longrightarrow V \text{ is a linear mapping;} \tag{3.9}$$

(b) A is continuous: for any $u, v \in V$

$$\begin{aligned}\langle Au, v\rangle &= a(u,v)\\ &\leq M\|u\|\,\|v\|.\end{aligned}$$

Taking $v = Au$ we obtain

$$\|Au\|^2 \leq M\|u\|\,\|Au\|,$$

which implies

$$\|Au\| \leq M\,\|u\|. \tag{3.10}$$

This expresses that A is a linear continuous operator from V into V with $\|A\|_{L(V,V)} \leq M$.

(c) A is coercive in the following sense: there exists some $\alpha > 0$ such that

$$\forall v \in V \qquad \langle Av, v\rangle \geq \alpha\|v\|^2. \tag{3.11}$$

To solve (3.8) we formulate it as a fixed point problem. Let λ be some strictly positive parameter. Clearly, to solve (3.8) is equivalent to finding $u \in V$ such that

$$u - \lambda(Au - f) = u. \tag{3.12}$$

In other words, we are looking for a fixed point $u \in V$ of the mapping $g_\lambda : V \longrightarrow V$ given by

$$g_\lambda(v) = v - \lambda(Av - f). \tag{3.13}$$

Let us prove that with λ adequately chosen, the mapping g_λ satisfies the condition of the Banach fixed point theorem, which we recall now.

Theorem 3.1.3 (Banach fixed point theorem—Picard iterative method). *Let (X, d) be a complete metric space and $g : X \longrightarrow X$ be a Lipschitz continuous mapping with a Lipschitz constant k strictly less than one, i.e.,*

$$\forall\, x, y \in X \quad d(g(x), g(y)) \leq k\, d(x,y).$$

Then, there exists a unique $\bar{x} \in X$ such that $g(\bar{x}) = \bar{x}$. Moreover, for any $x_0 \in X$, the sequence (x_n) starting from x_0 with $x_{n+1} = g(x_n)$ for all $n \in \mathbf{N}$ converges to $\bar{x}$ as n goes to $+\infty$.

Proof of Lax–Milgram theorem continued. Take $v_1, v_2 \in V$. Then

$$\begin{aligned}g_\lambda(v_2) - g_\lambda(v_1) &= [v_2 - \lambda(Av_2 - f)] - [v_1 - \lambda(Av_1 - f)]\\ &= (v_2 - v_1) - \lambda A(v_2 - v_1).\end{aligned}$$

Let us denote $v = v_2 - v_1$. So

$$\|g_\lambda(v_2) - g_\lambda(v_1)\| = \|v - \lambda Av\|.$$

To majorize this quantity, we consider its square and take advantage of the Hilbertian structure of V:

$$\begin{aligned}\|g_\lambda(v_2) - g_\lambda(v_1)\|^2 &= \|v - \lambda Av\|^2\\ &= \|v\|^2 - 2\lambda\langle Av, v\rangle + \lambda^2\|Av\|^2.\end{aligned}$$

By using (3.10) and (3.11) (note that we have assumed $\lambda > 0$), we obtain

$$||g_\lambda(v_2 - g_\lambda(v_1)||^2 \leq (1 - 2\lambda\alpha + \lambda^2 M^2)\, ||v||^2. \tag{3.14}$$

So, the question is to find some $\lambda > 0$ such that $1 - 2\lambda\alpha + \lambda^2 M^2 < 1$. Take $\bar{\lambda}$ for which the quantity $1 - 2\alpha\lambda + \lambda^2 M^2$ is minimal, that is, $\bar{\lambda} = \alpha / M^2$, in which case

$$1 - 2\bar{\lambda}\alpha + \bar{\lambda}^2 M^2 = 1 - \frac{\alpha^2}{M^2} < 1.$$

Hence

$$||g_{\bar{\lambda}}(v_2) - g_{\bar{\lambda}}(v_1)|| \leq \sqrt{1 - \frac{\alpha^2}{M^2}}\, ||v_2 - v_1||, \tag{3.15}$$

and $k_{\bar{\lambda}} = \sqrt{1 - \alpha^2 / M^2}$ is strictly less than one (note that $\alpha > 0$). So $g_{\bar{\lambda}}$ has a unique fixed point $\bar{u}$; equivalently, the equation $Au = f$ has a unique solution $\bar{u}$. □

Remark 3.1.2. (a) One of the main advantages of the proof above is that since it relies on the Banach fixed point theorem, it is a constructive proof.

(b) A second advantage is that it can be easily extended to nonlinear equations: solve

$$Au = f,$$

where $A : V \longrightarrow V$ satisfies

$$\exists M \geq 0 \text{ such that } \forall u, v \in V \quad ||Au - Av|| \leq M ||u - v||;$$

$$\exists \alpha > 0 \text{ such that } \forall u, v \in V \quad \langle Au - Av, u - v \rangle \geq \alpha ||u - v||^2.$$

(c) Another approach consists of proving that A is onto, that is, $R(A) = V$. To that end, one first establishes that

(i) $R(A)$ is closed: for this one can first notice that

$$\forall v \in V \quad \alpha ||v||^2 \leq \langle Av, v \rangle \leq ||Av|| \cdot ||v||$$

and hence

$$\alpha ||v|| \leq ||Av||.$$

If $Av_n \longrightarrow z$, we have

$$\alpha ||v_n - v_m|| \leq ||A(v_n - v_m)|| = ||Av_n - Av_m||$$

and (v_n) is a Cauchy sequence in V. Hence $v_n \longrightarrow v$ for some $v \in V$ and by continuity of A, $Av_n \longrightarrow Av$. Consequently $z = Av$.

(ii) $R(A)$ is dense: if $z \in R(A)^\perp$, then

$$\langle Av, z \rangle = 0 \quad \forall\, v \in V.$$

Take $v = z$ to conclude $z = 0$. Hence $R(A)^\perp = \{0\}$, that is, $\overline{R(A)} = V$.

3.1.2 ▪ The Galerkin method

A quite natural idea when considering an infinite dimensional (variational) problem is to approximate it by finite dimensional problems. This has important consequences both from the theoretical (existence, etc.) and the numerical point of view. In this section, we consider the situation corresponding to the Lax–Milgram theorem, and by using the Galerkin method we will both provide another proof of the existence of a solution and describe a corresponding approximation numerical schemes. We stress the fact that this type of finite dimensional approximation method is very flexible and can be applied (we will illustrate it further in various situations) to a large number of linear or nonlinear problems.

Definition 3.1.1. *Let V be a Banach space. A Galerkin approximation scheme is a sequence $(V_n)_{n\in\mathbf{N}}$ of finite dimensional subspaces of V such that for all $v \in V$, there exists some sequence $(v_n)_{n\in\mathbf{N}}$ with $v_n \in V_n$ for all $n \in \mathbf{N}$ and $(v_n)_{n\in\mathbf{N}}$ norm converging to v.*

This approximation property can be reformulated as

$$\forall v \in V \quad \lim_{n\to+\infty} \operatorname{dist}(v, V_n) = 0,$$

where $\operatorname{dist}(v, V_n) = \inf_{w\in V_n} \|v - w\|$.

Proposition 3.1.1. *Let V be a separable Banach space. Then one can construct a Galerkin approximation scheme (V_n) by the following method:*

(i) *take $(u_n)_{n\in\mathbf{N}}$ a countable dense subset of V (the separability of V just expresses that such set exists);*

(ii) *let $V_n = \operatorname{span}\{u_1, u_2, \dots, u_n\}$.*

Then $(V_n)_{n\in\mathbf{N}}$ is a Galerkin scheme.

Proof. Let $v \in V$. By (i), there exists a mapping $k \mapsto n(k)$ from $\mathbf{N}$ into $\mathbf{N}$ such that $\|u_{n(k)} - v\| \le \frac{1}{k}$ for all $k \in \mathbf{N}^*$. For k fixed, $u_{n(k)} \in V_{n(k)}$ and hence

$$\operatorname{dist}(v, V_{n(k)}) \le \frac{1}{k}.$$

Since $V_n \supset V_{n(k)}$ for $n \ge n(k)$, we obtain

$$\forall n \ge n(k) \quad \operatorname{dist}(v, V_n) \le \operatorname{dist}(v, V_{n(k)}) \le \frac{1}{k},$$

that is, $\operatorname{dist}(v, V_n) \longrightarrow 0$ as $n \to +\infty$. □

Remark 3.1.3. (a) The vectors $u_1, \dots, u_n$ need not to be linearly independent. By a classical linear algebra argument, one can replace in Proposition 3.1.1 the sequence $(u_n)_{n\in\mathbf{N}}$ by a sequence $(w_n)_{n\in\mathbf{N}}$, $w_n \in V$ made by linearly independent vectors.

(b) In Proposition 3.1.1 the sequence of subspaces $(V_n)_{n\in\mathbf{N}}$ satisfies

- $V_1 \subset V_2 \subset V_3 \subset \cdots \subset V_n \subset \cdots$ is an increasing sequence of finite dimensional subspaces;
- $\overline{U_{n\in\mathbf{N}} V_n} = V$.

The Galerkin approach to the Lax–Milgram theorem. We now suppose that V is a separable Hilbert space, $a : V \times V \longrightarrow \mathbf{R}$ is a bilinear, continuous, coercive form, $L : V \longrightarrow \mathbf{R}$ is a linear, continuous form.

We want to study the following problem: find $u \in V$ such that

$$a(u,v) = L(v) \quad \forall\, v \in V. \tag{3.16}$$

Since V is separable, by Proposition 3.1.1 there exists a Galerkin scheme $(V_n)_{n\in\mathbf{N}}$ with V_n increasing with $n \in \mathbf{N}$. Consider the approximated problems

$$\begin{cases} \text{find } u_n \in V_n \text{ such that} \\ a(u_n, v) = L(v) \quad \forall\, v \in V_n. \end{cases} \tag{3.17}$$

Problem (3.17) can be equivalently reformulated as

$$A_n u_n = f_n, \tag{3.18}$$

where A_n is the linear operator from $V_n \longrightarrow V_n$ such that

$$a(u,v) = \langle A_n u, v\rangle \qquad \forall\, u, v \in V_n$$

and

$$L(v) = \langle f_n, v\rangle \qquad \forall\, v \in V_n.$$

This is exactly the same argument as in the proof of the Lax–Milgram theorem except that now we work on finite dimensional spaces, which makes the existence of $(u_n)_{n\in\mathbf{N}}$ very easy: since $\ker A_n = 0$ (which follows from the coercivity of A_n), then A_n is onto. Note the basic difference with the infinite dimensional situation where such an argument is false!

The question now is to study the convergence of the sequence $(u_n)_{n\in\mathbf{N}}$. In (3.17) take $v = u_n$ so that

$$\alpha \|u_n\|^2 \le a(u_n, u_n) = L(u_n) \le \|L\|_* \|u_n\|$$

and

$$\|u_n\| \le \frac{\|L\|_*}{\alpha}. \tag{3.19}$$

The sequence $(u_n)_{n\in\mathbf{N}}$ is bounded and hence weakly relatively compact, that is, there exists a subsequence (u_{n_k}) and some $u \in V$ such that

$$\langle u_{n_k}, v\rangle \longrightarrow \langle u, v\rangle \ \forall v \in V. \tag{3.20}$$

We write $u_{n_k} \xrightarrow{w-V} u$. (See Theorem 2.4.3 with a direct independent proof of this result in separable Hilbert spaces.) Given $v \in V_m$, we have $V_{n_k} \supset V_m$ for k sufficiently large and hence

$$a(u_{n_k}, v) = L(v) \tag{3.21}$$

for all k sufficiently large. Then notice that

$$u_{n_k} \xrightarrow{w-V} u \Longrightarrow a(u_{n_k}, v) \longrightarrow a(u,v) \text{ as } k \longrightarrow +\infty.$$

This follows from the representation of the linear continuous form $u \mapsto a(u,v)$ on V, where $a(u,v) = \langle u, A^t v\rangle$ (A^t denotes the adjoint of A). Hence

$$u_{n_k} \xrightarrow{w-V} u \Longrightarrow a(u_{n_k}, v) = \langle u_{n_k}, A^t v\rangle \longrightarrow \langle u, A^t v\rangle = a(u,v).$$

So, when passing to the limit in (3.21), we obtain that, given $v \in V_m$,

$$a(u,v) = L(v).$$

Hence $a(u,v) = L(v)$ for every $v \in \bigcup_{m\in\mathbf{N}} V_m$. Since $\overline{\bigcup_{m\in\mathbf{N}} V_m} = V$ by continuity of $a(u,\cdot)$ and $L(\cdot)$ we finally infer

$$a(u,v) = L(v) \quad \forall\, v \in V.$$

Since u is the unique solution of this problem, by a classical compactness argument, the whole sequence $(u_n)_{n\in\mathbf{N}}$ weakly converges to u. Indeed, one can prove that the sequence $(u_n)_{n\in\mathbf{N}}$ norm converges to u. This is explained below with an explicit bound on $||u_n - u||$.

Proposition 3.1.2 (Cea lemma). *Let V be a separable Hilbert space and $(V_n)_{n\in\mathbf{N}}$ a Galerkin scheme. Suppose*

$$\begin{cases} a(u_n,v) = L(v) & \forall\, v \in V_n, \\ u_n \in V_n, \end{cases}$$

and

$$\begin{cases} a(u,v) = L(v) & \forall\, v \in V, \\ u \in V, \end{cases}$$

where a and L satisfy the assumptions of the Lax–Milgram theorem. Then

$$||u_n - u|| \le \frac{M}{\alpha}\, \mathrm{dist}(u, V_n).$$

PROOF. Subtracting (3.16) from (3.17), we obtain

$$a(u - u_n, v) = 0 \quad \forall v \in V_n,$$

and in particular

$$a(u - u_n, u_n) = 0.$$

It follows, for every $v \in V_n$,

$$\begin{aligned} a(u - u_n, u - u_n) &= a(u - u_n, u - v) + a(u - u_n, v - u_n) \\ &= a(u - u_n, u - v). \end{aligned}$$

Let us now use the continuity and coercivity property of a,

$$\alpha\, ||u - u_n||^2 \le M ||u - u_n|| \cdot ||u - v||,$$

to obtain

$$\alpha ||u - u_n|| \le M\, ||u - v|| \quad \forall\, v \in V_n.$$

Hence

$$||u - u_n|| \le \frac{M}{\alpha}\, \mathrm{dist}\,(u, V_n).$$

Since $(V_n)_{n\in\mathbf{N}}$ is a Galerkin scheme, $\mathrm{dist}(u, V_n) \longrightarrow 0$ as $n \longrightarrow +\infty$, and $(u_n)_{n\in\mathbf{N}}$ norm converges to u. □

3.2 ▪ Minimization problems: The topological approach

As a corollary of the Lax–Milgram theorem, we have obtained that if $a : V \times V \longrightarrow \mathbf{R}$ is bilinear, continuous, coercive, and symmetric, then, for any $L \in V^*$, there exists a unique solution to the minimization problem: find $u \in V$ such that

$$J(u) \leq J(v) \quad \forall v \in V,$$

where $J(v) = \frac{1}{2}\, a(v,v) - L(v)$.

Let us observe that $J : V \longrightarrow \mathbf{R}$ is convex, continuous, and coercive. By coercive, we mean that $\lim_{\|v\| \longrightarrow +\infty} J(v) = +\infty$, which follows easily from the inequality

$$J(v) \geq \frac{\alpha}{2}\, \|v\|^2 - \|L\|_* \cdot \|v\|.$$

Indeed, we will prove in Section 3.3.2 the following general result, which contains as a particular case the above situation (Theorem 3.3.4).

Let $J : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a real-extended valued function on a reflexive Banach space V, which is convex, lower semicontinuous, and coercive. Then there exists at least one $u \in V$ such that

$$J(u) \leq J(v) \quad \forall\, v \in V.$$

Before proving this theorem, in this section we will successively examine its basic ingredients. We will first justify the introduction of extended real-valued functions. Then, we will state the Weierstrass minimization theorem, which is purely topological, and in the process we will study the notions of lower semicontinuity, inf-compactness, and the interplay between inf-compactness, coercivity, and the role of the weak topology. So doing, we will be able to explain why convexity plays an important role in such questions.

3.2.1 ▪ Extended real-valued functions

A main reason for introducing extended real-valued functions is that they provide a natural and flexible modelization of minimization (or maximization) problems with constraints. Since in this chapter we consider minimization problems, we just need to consider functions

$$f : X \longrightarrow \mathbf{R} \cup \{+\infty\}.$$

On the other hand, if one considers maximization problems, one needs to introduce functions $f : X \longrightarrow \mathbf{R} \cup \{-\infty\}$. If maximization and minimization are both involved, just like in saddle value problems, one needs to work with functions $f : X \longrightarrow \bar{\mathbf{R}}$.

The (effective) *domain* of a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is the set

$$\operatorname{dom} f = \{x \in X \; : \; f(x) < +\infty\}.$$

The function f is said to be proper if $\operatorname{dom} f \neq \emptyset$.

Let us briefly justify the introduction of extended real-valued functions. Most minimization problems can be written as

$$\min\{f_0(x) \; : \; x \in C\}, \tag{3.22}$$

where $f_0 : X \longrightarrow \mathbf{R}$ is a real-valued function and $C \subset X$ is the set of constraints. In economics, C describes the available resources, the possible productions of a firm, or a set of decisions, and f_0 is the corresponding cost or economical criteria. In physics, the configurations x of the system are subject to constraints (unilateral or bilateral) and f_0, for example, is the corresponding energy.

A natural way to solve such problems is to approach them by penalization. For example, let us introduce a distance d on X and, for any positive real number k, consider the minimization problem

$$\min\{f_0(x) + k\, d(x,C) \;:\; x \in X\}, \tag{3.23}$$

where

$$d(x,C) = \inf\{d(x,y) \;:\; y \in C\} \tag{3.24}$$

is the distance function from x to C. Note that the penalization term is equal to zero if $x \in C$ (that is, if the constraint is fulfilled), and when $x \notin C$ (that is, if the constraint is violated) it takes larger and larger values which are increasing to $+\infty$ with k.

Let us also notice that the approximated problem (3.23) can be written as

$$\min\{f_k(x) \;:\; x \in X\},$$

where

$$f_k(x) = f_0(x) + k\, d(x,C)$$

is a real-valued function. Then the approximated problems are unconstrained problems, which makes the method interesting. As $k \to +\infty$, the sequence of functions $\{f_k \;:\; k \in \mathbf{N}\}$ increases to the function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, which is equal to

$$f(x) = \begin{cases} f_0(x) & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.25}$$

The function f is an extended real-valued function, so, if we want to treat in a unified way problems (3.22) and (3.23) we are naturally led to introduce extended real-valued functions. The minimization problem (3.22) can be equivalently formulated as

$$\min\{f(x) \;:\; x \in X\},$$

where f is given by (3.25). Note that in this formulation the constraint is equal to the domain of f. A particularly useful function in this unilateral framework is the *indicator function* δ_C of the set C:

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.26}$$

With this notation we have $f = f_0 + \delta_C$.

More generally, in variational analysis and optimization, one is often faced with expressions of the form

$$f = \sup_{i \in I} f_i$$

and one should notice that the class of extended real-valued functions is stable under such supremum operation. As a further illustration of these considerations, the convex duality theory establishes a one-to-one correspondence between a convex $C \subset X$ of a normed linear space and its support function

$$\sigma_C : X^* \longrightarrow \mathbf{R} \cup \{+\infty\},$$

which is defined by

$$\sigma_C(x^*) = \sup\{x^*(x) \;:\; x \in C\}$$

with X^* the topological dual of X. One cannot avoid the value $+\infty$ for σ_C as soon as the set C is not bounded (which is often the case!) and general duality statements must consider extended real-valued functions.

So, from now on, unless explicitly specified, we will consider functions $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ possibly taking the value $+\infty$. We adopt the conventions that $\lambda \times (+\infty) = +\infty$ if $\lambda > 0$ and $0 \times (+\infty) = 0$.

3.2.2 ▪ The interplay between functions and sets: The role of the epigraph

The analysis of unilateral problems (like minimization) naturally leads to the introduction of mathematical concepts which have a unilateral character. The classical approach of analysis does not provide the appropriate tools to deal with the mathematical objects and operations that are intrinsically unilateral (constraints, minimization). In the previous section, we justified the introduction of extended real-valued functions $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is the appropriate concept for dealing with minimization problems. Although in classical analysis the properties of the graph of a function play a fundamental role, in variational analysis it is the epigraph that will take over this role. The set

$$\operatorname{epi} f = \{(x,\lambda) \in X \times \mathbf{R} \;:\; \lambda \geq f(x)\} \tag{3.27}$$

is the epigraph of the function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$.

For any $\gamma \in \mathbf{R}$, the lower γ-level set of f is

$$lev_\gamma f = \{x \in X \;:\; f(x) \leq \gamma\}. \tag{3.28}$$

When considering the minimization problem of a given function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, the solution set is

$$\operatorname{argmin} f = \{\bar{x} \in X \;:\; f(\bar{x}) = \inf_X f(x)\}. \tag{3.29}$$

Note that $\operatorname{argmin} f$ can be possibly empty and

$$\operatorname{argmin} f = \bigcap_{\gamma > \inf_X f} lev_\gamma f. \tag{3.30}$$

Thus, to an extended real-valued function, we have associated many different sets—its epigraph, its lower level sets, its minimum set. Conversely, to a set we have associated extended real-valued functions, for example, the indicator function, the support function, the distance function. (This last one is real-valued if the set is nonempty.) In the following section, we describe how some basic topological properties of functions for minimization problems can be naturally formulated with the help of the attached geometrical sets, epigraphs, and lower level sets. Note that the lower level sets can be obtained by cutting operations on the epigraph:

$$lev_\gamma f \times \{\gamma\} = \operatorname{epi} f \cap (X \times \{\gamma\}).$$

Most basic operations in variational analysis can be naturally formulated with the help of the epigraphs. They give rise to the so-called epigraphical calculus; see [57], [68], [107]. The following result is valid for an arbitrary family of functions on an abstract space X.

Proposition 3.2.1. *Let X be an abstract space and $(f_i)_{i \in I}$ be a family of extended real-valued functions $f_i : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ indexed by an arbitrary set I. Then*

$$\operatorname{epi}(\sup_{i \in I} f_i) = \bigcap_{i \in I} \operatorname{epi} f_i,$$

$$\operatorname{epi}(\inf_{i \in I} f_i) = \bigcup_{i \in I} \operatorname{epi} f_i.$$

3.2.3 ▪ Lower semicontinuous functions

Let (X,τ) be a topological space. For any $x \in X$, we denote by $\mathcal{V}_\tau(x)$ the family of the neighborhoods of x for the topology τ.

We recall the classical definition of continuity for a function $f : (X,\tau) \longrightarrow \mathbf{R}$. The function f is said to be continuous at $x \in X$ for the topology τ if

$$\forall\, \varepsilon > 0 \quad \exists\, V_\varepsilon \in \mathcal{V}_\tau(x) \text{ such that } \forall\, y \in V_\varepsilon \quad |f(y) - f(x)| < \varepsilon.$$

This can be viewed as the conjunction of the two following properties:

1. $\forall \varepsilon > 0 \quad \exists\, V_\varepsilon \in \mathcal{V}_\tau(x)$ such that $\forall\, y \in V_\varepsilon \quad f(y) > f(x) - \varepsilon$;
2. $\forall \varepsilon > 0 \quad \exists\, W_\varepsilon \in \mathcal{V}_\tau(x)$ such that $\forall\, y \in W_\varepsilon \quad f(y) < f(x) + \varepsilon$.

Then, take $V_\varepsilon \cap W_\varepsilon$ which still belongs to $\mathcal{V}_\tau(x)$ to obtain the continuity result. These properties are called, respectively, the lower semicontinuity and the upper semicontinuity of f at x for the topology τ. To deal with possibly extended real-valued functions, definition 1 of lower semicontinuity has to be formulated slightly differently.

Definition 3.2.1. *Let (X,τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The function f is said to be τ-lower semicontinuous (τ-lsc) at x if*

$$\forall\, \lambda < f(x) \quad \exists V_\lambda \in \mathcal{V}_\tau(x) \quad \textit{such that } f(y) > \lambda \quad \forall\, y \in V_\lambda.$$

(We write V_λ to stress the dependence of the set V upon the choice of λ!)

If f is τ-lsc at every point of X, then f is said to be τ-lsc on X.

Proposition 3.2.2. *Let (X,τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function. The following statements are equivalent:*

(i) *f is τ-lsc;*

(ii) epi f *is closed in $X \times \mathbf{R}$ (where $X \times \mathbf{R}$ is equipped with the product topology of τ on X and of the usual topology on $\mathbf{R}$);*

(iii) *for all $\gamma \in \mathbf{R}$, $lev_\gamma f$ is closed in (X,τ);*

(iv) *for all $\gamma \in \mathbf{R}$, $\{x \in X \;:\; f(x) > \gamma\}$ is open in (X,τ);*

(v) *for all $x \in X$,* $\quad f(x) \leq \liminf_{y \to x} f(y) := \sup \inf_{V \in \mathcal{V}_\tau(x)\, y \in V} f(y)$.

PROOF. We are going to prove (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) $\Longrightarrow$ (v) $\Longrightarrow$ (i).

Assume that f is τ-lsc and prove that epi f is closed. Equivalently, let us prove that the complement of epi f in $X \times \mathbf{R}$ is open. Take $(x,\text{ł}) \notin \operatorname{epi} f$. By definition of epi f, $\text{ł} < f(x)$. Take $\text{ł} < \gamma < f$. Since f is τ-lsc, there exists $V_\gamma \in \mathcal{V}_\tau(x)$ such that

$$f(y) > \gamma \;\forall y \in V_\gamma.$$

Equivalently, $(y,\gamma) \notin \operatorname{epi} f$ for all $y \in V_\gamma$. It follows that

$$(V_\gamma \times]-\infty, \gamma[) \cap \operatorname{epi} f = \emptyset.$$

Noticing that $V_\gamma \times]-\infty,\gamma[$ is a neighborhood of (x,λ), the conclusion follows.

(ii) $\Longrightarrow$ (iii). The implication follows directly from the relation

$$lev_\gamma f \times \{\gamma\} = \operatorname{epi} f \cap (X \times \{\gamma\}).$$

Assuming that epi$f$ is closed, we infer that $lev_\gamma f \times \{\gamma\}$ is closed, and hence $lev_\gamma f$ is closed. (Note that $x \mapsto (x,\gamma)$ from X onto $X \times \{\gamma\}$ is a homeomorphism.)

(iii) $\Longrightarrow$ (iv) is obvious just by taking the complement of $lev_\gamma f$.

(iv) $\Longrightarrow$ (v). Let $\gamma < f(x)$. Since, by assumption, $\{y \in X \ : \ f(y) > \gamma\}$ is open, there exists some $V \in \mathcal{V}_\tau(x)$ such that $V \subset \{y \in X \ : \ f(y) > \gamma\}$. Equivalently,

$$\forall\, y \in V \quad f(y) > \gamma,$$

which implies

$$\inf_{y \in V} \; f(y) \geq \gamma.$$

Hence

$$\sup_{V \in \mathcal{V}_\tau(x)} \; \inf_{y \in V} f(y) \geq \gamma,$$

and this being true for any $\gamma < f(x)$, it follows that

$$\sup_{V \in \mathcal{V}_\tau(x)} \; \inf_{y \in V} f(y) \geq f(x),$$

that is,

$$f(x) \leq \liminf_{y \longrightarrow x} f(y).$$

(v) $\Longrightarrow$ (i). Let $\lambda < f(x)$. By assumption (v)

$$\lambda < \sup_{V \in \mathcal{V}_\tau(x)} \; \inf_{y \in V} f(y),$$

which implies the existence of some $V_\lambda \in \mathcal{V}_\tau(x)$ such that

$$\inf_{y \in V_\lambda} \; f(y) > \lambda,$$

i.e., $f(y) > \lambda$ for every $y \in V_\lambda$. This is exactly the lower semicontinuity property. □

The class of lower semicontinuous functions enjoys remarkable stability properties. Since closedness is preserved under arbitrary intersections and finite union of sets, we derive from Proposition 3.2.1 (epigraphical interpretation of sup and inf) and Proposition 3.2.2 (equivalence between f lsc and epi$f$ closed) the following important result.

Proposition 3.2.3. *Let (X,τ) be a topological space and $(f_i)_{i \in I}$, $f_i : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an arbitrary collection of τ-lsc functions. Then,* $\sup_{i \in I} f_i$ *is still τ-lsc. When I is a finite set of indices,* $\inf_{i \in I} f_i$ *is still τ-lsc.*

As a consequence, the supremum of a family of continuous functions is lower semicontinuous. One can prove that if (X,τ) is metrizable, the converse is true: if f is τ-lsc, there exists an increasing sequence $(f_n)_{n \in \mathbf{N}}$ of τ-continuous functions which is pointwise convergent to f. We will establish this important approximation result in Theorem 9.2.1 by using the epigraphical regularization. See also [86, Theorem 1.3.7].

Proposition 3.2.4. *Let $f, g : (X,\tau) \longrightarrow \mathbf{R}\cup\{+\infty\}$ be two lower semicontinuous functions. Then $f+g$ is still lower semicontinuous.*

PROOF. Take $x_\nu \longrightarrow x$ a τ-converging net. Then

$$\begin{aligned}\liminf_\nu (f+g)(x_\nu) &= \liminf_\nu [f(x_\nu)+g(x_\nu)]\\ &\geq \liminf_\nu f(x_\nu) + \liminf_\nu g(x_\nu)\\ &\geq f(x)+g(x). \qquad \square\end{aligned}$$

3.2.4 ▪ The lower closure of a function and the relaxation problem

In some important situations, the function f to minimize fails to be lower semicontinuous for a topology τ which makes a minimizing sequence τ-relatively compact (see Section 3.2.5). In that case, the analysis of the behavior of the minimizing sequences requires the introduction of the lower closure of f and leads to the introduction of the relaxed problem.

Definition 3.2.2. *Given (X,τ) a topological space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$, the τ-lower envelope of f is defined as*

$$cl_\tau f = \sup\{g : X \to \mathbf{R}\cup\{+\infty\} \ : \ g \ \tau\text{-}lsc, \ g \leq f\}.$$

Proposition 3.2.5. *Let (X,τ) be a topological space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function. Then $cl_\tau f$ is τ-lsc; it is the greatest τ-lsc function which minorizes f. We have the following properties:*

(a) epi$\,(cl_\tau f)$ *is the closure of* epi$\,f$ *in $X\times\mathbf{R}$ equipped with the product topology of τ with the usual topology of* $\mathbf{R}$*:*

$$\operatorname{epi}(cl_\tau f) = cl(\operatorname{epi} f);$$

(b) $cl_\tau f = \sup_{V\in\mathcal{V}_\tau(x)} \inf_{y\in V} f(y) = \liminf_{y\longrightarrow x} f(y)$*;*

(c) *f is τ-lsc at x iff $f(x)\leq cl_\tau f(x)$;*

(d) *f is τ-lsc at x iff $f(x)= cl_\tau f(x)$.*

PROOF. Let us first notice that a set C in $X\times\mathbf{R}$ is an epigraph iff

(i) C recedes in the vertical direction:

$$(x,\text{ł})\in C \text{ and } \mu > \text{ł} \Longrightarrow (x,\mu)\in C;$$

(ii) C is vertically closed:

$$\text{for every } x\in X \text{ the set } \{\text{ł}\in\mathbf{R} \ : \ (x,\text{ł})\in C\} \text{ is closed.}$$

(a) This implies that the closure of an epigraph is an epigraph. Set $cl(\operatorname{epi} f) = \operatorname{epi} g$. Since epi g is closed, this implies that g is τ-lsc. Moreover, since epi $g \supset$ epi$f$, we have $g\leq f$. Hence g is a τ-lsc minorant of f. We claim that g is the lower envelope of f, that is, g is the greatest of such τ-lsc minorants of f.

Take $h\leq f$ and h τ-lsc. Hence

$$\operatorname{epi} h \supset \operatorname{epi} f,$$

which implies

$$cl(\operatorname{epi} h) = \operatorname{epi} h \supset cl(\operatorname{epi} f) = \operatorname{epi} g.$$

Hence, $\operatorname{epi} h \supset \operatorname{epi} g$ and $h \leq g$. So

$$g = \sup\{h : X \to \mathbf{R} \cup \{+\infty\} \,:\, h\ \tau\text{-}lsc,\ h \leq f\},$$

that is, $g = cl_\tau f$. We have proved that

$$cl(\operatorname{epi} f) = \operatorname{epi} g = \operatorname{epi}(cl_\tau f).$$

(b) Since $cl_\tau f$ is τ-lsc, it follows from Proposition 3.2.2(v) that

$$(cl_\tau f)(x) \leq \liminf_{y \longrightarrow x} (cl_\tau f)(y),$$

and since $cl_\tau f \leq f$,

$$(cl_\tau f)(x) \leq \liminf_{y \longrightarrow x} f(y) \quad \forall\, x \in X. \tag{3.31}$$

Then notice that the function $h(x) = \liminf_{y \longrightarrow x} f(y)$ is less than or equal to f and is τ-lsc. To verify this last point, take

$$\lambda < \liminf_{y \longrightarrow x} f(y) = \sup_{V \in \mathscr{V}_\tau(x)} \inf_{y \in V} f(y).$$

Then there exists some open set $V_\lambda \in \mathscr{V}_\tau(x)$ such that $\inf_{y \in V_\lambda} f(y) > \lambda$. It follows that for all $\xi \in V_\lambda$, $V_\lambda \in \mathscr{V}_\tau(\xi)$ and hence

$$\sup_{V \in \mathscr{V}_\tau(\xi)} \inf_{y \in V} f(y) > \lambda,$$

that is,

$$\liminf_{y \longrightarrow \xi} f(y) > \lambda.$$

Thus, $x \mapsto \liminf_{y \longrightarrow x} f(x)$ is τ-lsc and minorizes f. By definition of $cl_\tau f$ we have

$$\liminf_{y \longrightarrow x} f(y) \leq (cl_\tau f)(x). \tag{3.32}$$

Then compare (3.31) and (3.32) to obtain

$$\forall x \in X \quad (cl_\tau f)(x) = \liminf_{y \longrightarrow x} f(y).$$

(c) Proposition 3.2.2 expresses that f is τ-lsc at x iff

$$f(x) \leq \liminf_{y \longrightarrow x} f(y).$$

This is equivalent to saying that

$$f(x) \leq cl_\tau f(x),$$

and since $cl_\tau f \leq f$ is always true, it is equivalent to $f(x) = cl_\tau f(x)$. □

We have the following "sequential" formulation of $cl_\tau f$.

Proposition 3.2.6. *Let (X,τ) be a topological space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$. Then, for any $x \in X$,*

$$\begin{aligned}(cl_\tau f)(x) &= \liminf_{y\longrightarrow x} f(y)\\ &= \min\{\liminf_{\nu} f(x_\nu) \ :\ x_\nu \text{ is a net }, x_\nu \longrightarrow x\}.\end{aligned}$$

When (X,τ) is metrizable

$$\begin{aligned}(cl_\tau f)(x) &= \liminf_{y\longrightarrow x} f(y)\\ &= \min\{\liminf_{n} f(x_n) \ :\ (x_n) \text{ sequence, } x_n \longrightarrow x\}.\end{aligned}$$

Corollary 3.2.1. *Let (X,τ) be a metrizable space and let $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ be an extended real-valued function. Then f is τ-lsc at $x \in X$ iff*

$$\forall x_n \longrightarrow x \quad f(x) \le \liminf_n f(x_n).$$

Proof of Proposition 3.2.6. We give for simplicity the proof only in the metrizable case. We have

$$\liminf_{y\longrightarrow x} f(y) = \sup_{\varepsilon>0}\ \inf_{y\in B_\tau(x,\varepsilon)} f(y),$$

where $B_\tau(x,\varepsilon) = \{y \in X \ :\ d_\tau(y,x) < \varepsilon\}$, d_τ being a distance inducing the topology τ. Then, for any $x_n \longrightarrow x$, for any $\varepsilon > 0$, x_n belongs to $B_\tau(x,\varepsilon)$ for n sufficiently large. Hence

$$\inf_{y\in B_\tau(x,\varepsilon)} f(y) \le f(x_n) \quad \forall n \ge N(\varepsilon).$$

Passing to the limit as $n \to +\infty$ gives

$$\inf_{y\in B_\tau(x,\varepsilon)} f(y) \le \liminf_n f(x_n).$$

This being true for any $\varepsilon > 0$ and any $x_n \longrightarrow x$, leads to the inequality

$$\liminf_{y\longrightarrow x} f(y) \le \inf\{\liminf_n f(x_n) \ :\ x_n \longrightarrow x\}. \tag{3.33}$$

On the other hand, for each $n \in \mathbf{N}$, there exists some $x_n \in B_\tau(x,1/n)$ such that

$$\begin{aligned}\inf_{y\in B_\tau(x,1/n)} f(y) \ge f(x_n) - \frac{1}{n} \quad &\text{if} \inf_{y\in B_\tau(x,1/n)} f(y) > -\infty\\ -n \ge f(x_n) \quad &\text{if} \inf_{y\in B_\tau(x,1/n)} f(y) = -\infty.\end{aligned}$$

In both cases,

$$\begin{aligned}\liminf_{y\longrightarrow x} f(y) = \lim_n \inf_{y\in B_\tau(x,1/n)} f(y) &\ge \limsup_n f(x_n)\\ &\ge \liminf_n f(x_n).\end{aligned} \tag{3.34}$$

Then compare (3.33) and (3.34) to obtain

$$\liminf_{y \longrightarrow x} f(y) = \min\{\liminf_n f(x_n) \ : \ x_n \longrightarrow x\}$$
$$= \min\{\limsup_n f(x_n) \ : \ x_n \longrightarrow x\}. \qquad \square$$

Proposition 3.2.7. *Let $f : (X, \tau) \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. Then*

$$\inf_X f = \inf_X cl_\tau f.$$

More generally, for any τ-open subset G of X

$$\inf_G f = \inf_G cl_\tau f.$$

Moreover,

$$\operatorname{argmin} f \subset \operatorname{argmin} cl_\tau f.$$

PROOF. (a) Since $f \geq cl_\tau f$, we just need to prove that

$$\inf_G cl_\tau f \geq \inf_G f.$$

For any $x \in G$, since G is τ-open, we have $G \in \mathcal{V}_\tau(x)$. By Proposition 3.2.5,

$$(cl_\tau f)(x) = \sup_{V \in \mathcal{V}_\tau(x)} \inf_{y \in V} f(y).$$

Hence, for every $x \in G$

$$(cl_\tau f)(x) \geq \inf_{y \in G} f(y).$$

This being true for any $x \in G$ gives

$$\inf_G (cl_\tau f) \geq \inf_G f.$$

(b) Let $x \in \operatorname{argmin} f$. We have

$$cl_\tau f(x) \leq f(x) \leq \inf_X f = \inf_X cl_\tau f,$$

which implies $x \in \operatorname{argmin} cl_\tau f$. $\square$

Remark 3.2.1. The function $cl_\tau f$, which we call the lower envelope of f, is often called the lower semicontinuous regularization of f or the *relaxed function* or the τ-closure. The problem $\min\{cl_\tau f(x) \ : \ x \in X\}$ is called the *relaxed problem*.

3.2.5 ▪ Inf-compactness functions, coercivity

Besides lower semicontinuity, the second basic ingredient in minimization problems is the inf-compactness property.

Definition 3.2.3. *Let (X, τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. The function f is said to be τ-inf-compact if for any $\gamma \in \mathbf{R}$*

$$lev_\gamma f = \{x \in X \ : \ f(x) \leq \gamma\}$$

is relatively compact in X for the topology τ.

When f is τ-lsc, its lower level sets are closed for τ, and τ-compactness is equivalent to saying that the lower level sets of f are τ-compact.

Definition 3.2.4. *Let X be a normed linear space. A function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is said to be coercive if* $\lim_{||x|| \longrightarrow +\infty} f(x) = +\infty$.

The relation between the two concepts is made clear in the following.

Proposition 3.2.8. *Let X be a normed space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The following conditions are equivalent:*

(i) *f is coercive;*

(ii) *for any $\gamma \in \mathbf{R}$, $lev_\gamma f$ is bounded.*

PROOF. (i) $\Longrightarrow$ (ii). Assume f is coercive. If, for some $\gamma_0 \in \mathbf{R}$, $lev_{\gamma_0} f$ is not bounded, then there exists a sequence $(x_n)_{n\in\mathbf{N}}$ such that $||x_n|| \longrightarrow +\infty$ as n goes to $+\infty$ and $f(x_n) \le \gamma_0$ for all $n \in \mathbf{N}$. But f coercive implies that $f(x_n) \longrightarrow +\infty$, a clear contradiction.

(ii) $\Longrightarrow$ (i). Assume that for any $\gamma \in \mathbf{R}$, $lev_\gamma f$ is bounded. If f is not coercive, we can construct a sequence $(x_n)_{n\in\mathbf{N}}$ such that $||x_n|| \longrightarrow +\infty$ as n goes to $+\infty$ and such that $f(x_n) \le \gamma_0$ for some $\gamma_0 \in \mathbf{R}$. This contradicts the fact that $lev_{\gamma_0} f$ is bounded. □

Let us recall the Riesz theorem: the bounded sets in a normed space are relatively compact iff the space has a finite dimension.

Corollary 3.2.2. *Let $X = \mathbf{R}^n$ equipped with the usual topology and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The following conditions are equivalent:*

(i) *f is coercive;*

(ii) *f is* inf-*compact.*

In infinite dimensional spaces, the topologies which are directly related to coercivity are the weak topologies (see Section 2.4).

3.2.6 ▪ Topological minimization theorems

In this section, unless otherwise specified (X, τ) is a general topological space.

Theorem 3.2.1. *Let (X, τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function which is τ-lsc and τ-inf compact. Then* $\inf_X f > -\infty$ *and there exists some $\bar{x} \in X$ which minimizes f on X:*

$$f(\bar{x}) \le f(x) \quad \forall\, x \in X.$$

Because of the importance of this theorem, which is often referred to as the Weierstrass theorem, we give two different proofs of it, each of independent interest. Without any restriction we may assume that f is proper, that is, $f \not\equiv +\infty$.

FIRST PROOF. We want to prove that $\operatorname{arg\,min} f \neq \emptyset$. We use formula (3.30), which relates $\operatorname{arg\,min} f$ to the lower level sets of f.

$$\begin{aligned} \operatorname{arg\,min} f &= \bigcap_{\gamma > \inf_X f} lev_\gamma f \\ &= \bigcap_{\gamma_0 > \gamma > \inf_X f} lev_\gamma f, \end{aligned}$$

where $\gamma_0 \in \mathbf{R}$ is taken arbitrary with $\gamma_0 > \inf_X f$. This comes from the fact that the sets $lev_\gamma f$ are decreasing with γ. The τ-lower semicontinuity of f implies that the sets $lev_\gamma f$ are closed for the topology τ. Moreover, for $\gamma_0 > \gamma > \inf_X f$ the sets $lev_\gamma f$ are nonempty and contained in $lev_{\gamma_0} f$ which is compact by the τ-inf compact property of f.

Therefore, we have a family $\{lev_\gamma f \ : \ \gamma_0 < \gamma < \inf_X f\}$ of nonempty closed subsets, contained in a fixed compact set, and which is decreasing with γ. For any finite subfamily $\{lev_{\gamma_i} f \ : \ i = 1, 2, \dots, m\}$

$$\bigcap_{i=1,\dots,m} lev_{\gamma_i} f = lev_\gamma f$$

with $\gamma = \inf\{\gamma_1, \dots, \gamma_m\} > \inf_X f$, and hence

$$\bigcap_{i=1,\dots,m} lev_{\gamma_i} f \neq \emptyset.$$

From the finite intersection property (which characterizes topological compact sets; it is obtained from the Heine–Borel property just by passing to the complement, and so replacing open sets by closed sets), we conclude that $\bigcap_{\gamma_0 > \gamma > \inf_X f} lev_\gamma f \neq \emptyset$.

SECOND PROOF. The proof we present now illustrates the direct method in the calculus of variations. It is the proof initiated by Hilbert and further developed by Tonelli, which first introduces a minimizing sequence.

Let us observe that given a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, one can always construct a minimizing sequence, that is, a sequence $(x_n)_{n \in \mathbf{N}}$ such that $f(x_n) \longrightarrow \inf_X f$ as $n \longrightarrow +\infty$. To do so, we just rely on the definition of the infimum of a family of real numbers:

if $\inf_X f > -\infty$, take $\inf_X f \leq f(x_n) \leq \inf_X f + 1/n$;
if $\inf_X f = -\infty$, take $f(x_n) \leq -n$.

Since f is proper, $\inf_X f < +\infty$ and for $n \geq 1$

$$\begin{aligned} f(x_n) &\leq \max\{\inf_X f + 1/n, -n\} \\ &\leq \max\{\inf_X f + 1, -1\} := \gamma_0. \end{aligned}$$

Note that $\gamma_0 > \inf_X f$, which implies $lev_{\gamma_0} f \neq \emptyset$ and

$$x_n \in lev_{\gamma_0} f \quad \forall\, n \geq 1.$$

Thus, the sequence $(x_n)_{n \in \mathbf{N}}$ is trapped in a lower level set of f which is compact for the topology τ (f is τ-inf compact). We follow the argument and assume that τ is metrizable. In the general topological case, one has to replace sequences by nets. So, we can extract a subsequence τ-converging to some $\bar{x} \in X$,

$$x_{n_k} \longrightarrow \bar{x}.$$

We have

$$\lim_k f(x_{n_k}) = \lim_n f(x_n) = \inf_X f.$$

By the τ-lower semicontinuity of f,

$$f(\bar{x}) \leq \lim_k f(x_{n_k}).$$

Hence,

$$f(\bar{x}) \leq \inf_X f,$$

which says both that $\inf_X f > -\infty$ since $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ and

$$f(\bar{x}) \leq f(x) \quad \forall\, x \in X,$$

and the proof is complete. □

Remark 3.2.2. Indeed, the inf-compactness assumption can be slightly weakened by noticing that in the proof of Theorem 3.2.1, we just need to know that some lower level set of f is relatively compact. Notice that it is equivalent to assume that $lev_{\gamma_0} f$ is relatively compact or to assume that $lev_\gamma f$ is relatively compact for all $\gamma \leq \gamma_0$. So, we can formulate the following result.

Theorem 3.2.2. *Let (X, τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is τ-lsc and such that for some $\gamma_0 \in \mathbf{R}$, $lev_{\gamma_0} f$ is τ-compact.*

Then $\inf_X f > -\infty$ and there exists some $\bar{x} \in X$ which minimizes f on X:

$$f(\bar{x}) \leq f(x) \quad \forall\, x \in X.$$

To illustrate the difference between Theorems 3.2.1 and 3.2.2, take $X = \mathbf{R}$ and $f(x) = \frac{x^2}{1+x^2}$. Then $lev_\gamma f$ is compact for $\gamma < 1$, but $lev_1 f = \mathbf{R}$. Thus we can apply Theorem 3.2.2 to conclude the existence of a minimizer (which is zero!), but Theorem 3.2.1 does not apply!

Corollary 3.2.3. *Let (X, τ) be a topological space and assume that $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is τ-lsc. Then, for any compact subset K of (X, τ), there exists some $\bar{x} \in K$ such that*

$$f(\bar{x}) \leq f(x) \quad \forall x \in K.$$

Proof. Take $g := f + \delta_K$. Since K is closed, δ_K is lower semicontinuous. So, g as a sum of two lower semicontinuous functions is still lower semicontinuous. The sublevel sets of g are contained in K, so g is τ-inf compact. Therefore, by Theorem 3.2.1 there exists some $\bar{x} \in X$ such that $g(\bar{x}) \leq g(x)$ for every $x \in X$, that is,

$$\begin{cases} f(\bar{x}) \leq f(x) & \forall\, x \in K, \\ \bar{x} \in K, \end{cases}$$

which completes the proof. □

The above statement is the unilateral version of the classical theorem which says that a continuous function achieves on any compact its minimum value and maximum value. If one is concerned only with the minimization problem, one just needs to consider lower semicontinuous functions.

Corollary 3.2.4. *Take $X = \mathbf{R}^N$ with the usual topology. Take $f : \mathbf{R}^N \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is lower semicontinuous and coercive. Then, there exists some $\bar{x} \in \mathbf{R}^N$ such that*

$$f(\bar{x}) \leq f(x) \quad \forall x \in \mathbf{R}^N.$$

Proof. Since $f : \mathbf{R}^N \longrightarrow \mathbf{R} \cup \{+\infty\}$ is coercive, it is inf-compact for the usual topology (Corollary 3.2.2). This combined with the lower semicontinuity of f implies the existence of a minimizer. □

Comments on the direct methods of the calculus of variations. Theorems 3.2.1 and 3.2.2 provide both a general existence result for the global minimization problem of an extended real-valued function and a method for solving such problems, originally introduced by Hilbert and further developed by Tonelli: when considering a minimization problem for a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$

$$\min\{f(x) \,:\, x \in X\},$$

one first constructs a minimizing sequence, which is a sequence $(x_n)_{n\in\mathbf{N}}$ such that

$$f(x_n) \longrightarrow \inf_X f \quad \text{as } n \longrightarrow +\infty.$$

This is always possible; at this point, we don't need any structure on X. Then, one has to establish that the sequence $(x_n)_{n\in\mathbf{N}}$ is relatively compact for some topology τ on X, and this is how the topology τ appears. This usually comes from some estimations on $(x_n)_{n\in\mathbf{N}}$ which follow from a coercivity property of f with respect to some norm on X. Then, one may use weak topologies or some compact embeddings to find the topology τ.

We stress that there is a great flexibility in this method. One may consider special minimizing sequences enjoying compactness properties which are not shared by the whole lower level sets. We will return to this important point. We just say here that it is the skill of the mathematician to find a minimizing sequence which is relatively compact for a topology τ which is as strong as possible.

Indeed, and this is the second point of the direct methods, one then has to verify that f is τ-lsc. Clearly, the stronger the topology τ, the easier it is to verify the lower semicontinuity property.

As a general rule, inf-compactness and lower semicontinuity are two properties which are antagonist: if $\tau_1 > \tau_2$, then

$$f \ \tau_1 \quad \text{inf-compact} \Longrightarrow f \ \tau_2 \quad \text{inf-compact},$$

while

$$f \ \tau_2\text{—lsc} \Longrightarrow f \ \tau_1\text{-lsc}.$$

Thus, a balance with respect to these two properties determines the choice of a "good" topology τ (if it exists!).

As we will see, in some important situations, the function f fails to be τ-lsc and there is no solution to the minimization problem of f. In such situations, it is still interesting to understand the behavior of the minimizing sequences. The following "relaxation result" gives a first general answer to this question.

Theorem 3.2.3. *Let (X,τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. Let $(x_n)_{n\in\mathbf{N}}$ be a minimizing sequence for f, and suppose that a subsequence x_{n_k} τ-converges to some $\bar{x} \in X$. Then*

$$(cl_\tau f)(\bar{x}) \le (cl_\tau f)(x) \quad \forall\, x \in X,$$

that is, $\bar{x}$ is a minimum point for $cl_\tau f$.

PROOF. Since $(x_n)_{n\in\mathbf{N}}$ is a minimizing sequence,

$$\lim_k f(x_{n_k}) = \lim_n f(x_n) = \inf_X f.$$

By Proposition 3.2.6, since $\bar{x} = \tau - \lim x_{n_k}$,

$$cl_\tau f(\bar{x}) \leq \lim_k f(x_{n_k}).$$

By Proposition 3.2.7,

$$\inf_X f = \inf_X cl_\tau f.$$

Hence

$$cl_\tau f(\bar{x}) \leq \lim f(x_{n_k}) = \lim f(x_n) = \inf_X f = \inf_X cl_\tau f,$$

that is,

$$(cl_\tau f)(\bar{x}) \leq (cl_\tau f)(x) \quad \forall x \in X. \qquad \square$$

We say that $\min\{(cl_\tau f)(x) : x \in X\}$ is the *relaxed problem* of the initial minimization problem of f over X.

3.2.7 ▪ Weak topologies and minimization of weakly lower semicontinuous functions

Until now, the basic ingredients used in the direct approach to minimization problems have been purely topological notions. We now assume that the underlying space (X,τ) is a vector space. To stress that fact, we denote it by V (like vector) and assume that V is a normed space, the norm of $v \in V$ being denoted by $||v||_V$ or $||v||$ when no confusion is possible.

The consideration of coercive functions $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ leads naturally to study of the topological properties of the bounded subsets of V, and this is a basic reason for studying weak topologies on topological vector spaces.

We recall from Theorems 2.4.2 and 2.4.3 the following result.

Theorem 3.2.4. *In a reflexive Banach space V, the bounded sets are weakly relatively compact. Moreover, from any bounded sequence $(u_n)_{n\in\mathbf{N}}$ in V one can extract a weakly convergent subsequence.*

As a direct application of Proposition 3.2.8 and of the previous compactness result, we obtain that a coercive function on a reflexive Banach space is weakly inf-compact. By using the Weierstrass minimization Theorem 3.2.1 we obtain the following existence result.

Theorem 3.2.5. *Let V be a reflexive Banach space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is coercive and weakly lower semicontinuous. Then there exists some $u \in V$ such that*

$$f(u) \leq f(v) \quad \forall v \in V.$$

The question that now naturally arises is to describe the class of functions which are weakly lower semicontinuous. This is where convexity plays a central role.

3.3 ▪ Convex minimization theorems

Let us first recall some definitions and elementary properties of extended real-valued convex functions.

3.3.1 ▪ Extended real-valued convex functions and weak lower semicontinuity

Definition 3.3.1. *Let V be a linear space and let $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then, f is said to be convex if for each $u, v \in V$ and each $\lambda \in [0,1]$ we have*

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v).$$

Proposition 3.3.1. *Let V be a linear space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then, f is convex iff its epigraph is a convex subset of $V \times \mathbf{R}$.*

PROOF. Let us first assume that f is convex. Fix (u,α) and (v,β) in $\operatorname{epi} f$ and $\lambda \in [0,1]$. Since $\alpha \geq f(u)$ and $\beta \geq f(v)$, then $f(u)$ and $f(v)$ are finite and we have

$$\begin{aligned}\lambda\alpha + (1-\lambda)\beta &\geq \lambda f(u) + (1-\lambda)f(v)\\ &\geq f(\lambda u + (1-\lambda)v).\end{aligned}$$

This is equivalent to saying that $(\lambda u + (1-\lambda)v, \lambda\alpha + (1-\lambda)\beta) \in \operatorname{epi} f$, i.e.,

$$\lambda(u,\alpha) + (1-\lambda)(v,\beta) \in \operatorname{epi} f,$$

and so $\operatorname{epi} f$ is convex.

Conversely, let us assume that $\operatorname{epi} f$ is convex. Fix $\lambda \in [0,1]$. If either $f(u) = +\infty$ or $f(v) = +\infty$, since $0 \times (+\infty) = 0$, the inequality is clearly valid. So, let us assume that $f(u) < +\infty$ and $f(v) < +\infty$. The two points $(u, f(u))$ and $(v, f(v))$ are in $\operatorname{epi} f$ and so is the segment joining this two points. In particular,

$$\lambda(u, f(u)) + (1-\lambda)(v, f(v)) \in \operatorname{epi} f,$$

which is equivalent to saying that $f(\lambda u + (1-\text{ł})v) \leq \lambda f(u) + (1-\lambda)f(v)$. □

So, it is equivalent to study convex sets or convex functions. Let us now recall the geometrical version of the Hahn–Banach theorem, which plays a basic role in convex analysis (see Chapter 9).

Theorem 3.3.1 (Hahn–Banach separation theorem). *Let $(V, ||\cdot||)$ be a normed linear space and suppose that C is a nonempty closed convex subset of V. Then, each point $u \notin C$ can be strongly separated from C by a closed hyperplane, which means*

$$\exists u^* \in V^*,\ \exists \alpha \in \mathbf{R} \ \textit{such that}\ \forall v \in C \quad u^*(v) \leq \alpha \ \textit{and}\ u^*(u) > \alpha.$$

This is equivalent to saying that C is contained in the closed half-space $\mathscr{H}_{\alpha,u^*} = \{v \in V : u^*(v) \leq \alpha\}$, whereas u is in its complement.

Corollary 3.3.1. *Let $(V, ||\cdot||)$ be a normed linear space and let C be a nonempty closed convex subset of V. Then C is equal to the intersection of the closed half-spaces that contain it.*

Theorem 3.3.1 and Corollary 3.3.1 have important consequences with respect to topological (closedness) properties of convex sets: let us notice that by definition of the weak topology on V, any linear strongly continuous form is continuous for the weak topology, which implies that any closed half-space

$$\mathscr{H}_{\alpha,u^*} = \{v \in V : u^*(v) \leq \alpha\}$$

is closed for the weak topology. So, any closed convex set, which is equal to an intersection of closed half-spaces, is closed for the weak topology. Since for an arbitrary set, the reverse implication "closed for the weak topology" $\Longrightarrow$ "closed for the strong topology" is always true, we finally obtain the following result.

Theorem 3.3.2. *Let $(V, \|\cdot\|)$ be a normed linear space and C a nonempty convex subset of V. Then, the following statements are equivalent:*

(i) *C is closed for the norm topology of V;*

(ii) *C is closed for the weak topology of V.*

When translating the above theorem from sets to functions via the correspondence $f \longrightarrow \operatorname{epi} f$ and recalling that

$$\begin{aligned} &f \text{ convex } \Longleftrightarrow \operatorname{epi} f \text{ convex in } X \times \mathbf{R}, \\ &f \ \tau\text{-}lsc \Longleftrightarrow \operatorname{epi} f \text{ closed in } X \times \mathbf{R}, \end{aligned}$$

we obtain the following result.

Theorem 3.3.3. *Let V be a normed linear space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ a convex proper function. The following statements are equivalent:*

(i) *f is lower semicontinuous for the norm topology on V;*

(ii) *f is lower semicontinuous for the weak topology on V.*

Of course it is the implication (i) $\Longrightarrow$ (ii) which is important. It tells us that a convex lower semicontinuous function is automatically weakly lower semicontinuous.

As a particular case, a convex continuous function is weakly lower semicontinuous.

3.3.2 ▪ Convex minimization in reflexive Banach spaces

In this section $(V, \|\cdot\|)$ is a reflexive Banach space. We can now state the following important result.

Theorem 3.3.4. *Let $(V, \|\cdot\|)$ be a reflexive Banach space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ a convex, lower semicontinuous, and coercive function. Then there exists $u \in V$ which minimizes f on V:*

$$f(u) \leq f(v) \quad \forall\, v \in V.$$

FIRST PROOF. Since f is coercive, its lower level sets are bounded in V and hence weakly relatively compact. So f is weakly inf-compact. Since f is convex lower semicontinuous, it is weakly lower semicontinuous. Then apply the Weierstrass minimization Theorem 3.2.1 to f with τ equal to the weak topology of V.

SECOND PROOF. We use the direct methods of the calculus of variations. Take $(u_n)_{n \in \mathbf{N}}$, a minimizing sequence for f, that is, $\lim_n f(u_n) = \inf_V f$. For n sufficiently large, $(u_n)_{n \in \mathbf{N}}$ remains in a fixed sublevel set of f, which is bounded by coercivity of f. Because the space V is reflexive, one can extract a weakly convergent subsequence $u_{n_k} \rightharpoonup u$. We have

$$\lim_k f(u_{n_k}) = \lim_n f(u_n) = \inf_V f.$$

The function f is convex and lower semicontinuous, so it is weakly lower semicontinuous and

$$f(u) \le \liminf_{k} f(u_{n_k}).$$

It follows that

$$f(u) \le \liminf_{k} f(u_{n_k}) = \lim_{n} f(u_n) = \inf_{V} f,$$

that is, $f(u) \le f(v)\ \forall v \in V$. □

Remark 3.3.1. Concerning the question of uniqueness, we need to recall the notion of strict convexity: the function $f \longrightarrow \mathbf{R} \cup \{+\infty\}$ is said to be strictly convex if

$$\forall u \neq v, \quad \forall \lambda \in]0,1[\quad f(\lambda u + (1-\lambda)v) < \lambda f(u) + (1-\lambda) f(v).$$

It is easily seen that the conclusion of Proposition 2.3.3 still holds for functions with values in $\mathbf{R} \cup \{+\infty\}$.

Example 3.3.1. Take for V a Hilbert space with norm $||\cdot||^2 = \langle \cdot,\cdot \rangle$. Then $||\cdot||^2$ is strictly convex: indeed, for $v_1, v_2 \in V$, $\lambda \in]0,1[$ we have

$$\begin{aligned}
&||\lambda v_1 + (1-\lambda)v_2||^2 - \lambda ||v_1||^2 - (1-\lambda)||v_1||^2 \\
&\quad = -\lambda(1-\lambda)\big[||v_2||^2 + ||v_1||^2 - 2\langle v_1, v_2\rangle\big] \\
&\quad = -\lambda(1-\lambda)||v_1 - v_2||^2 \\
&\quad \le 0.
\end{aligned}$$

The above inequality becomes an equality iff $v_1 = v_2$.

As an application of the results above, we consider the problem of the best approximation in Hilbert spaces.

Theorem 3.3.5. *Let $(V, ||\cdot||)$, be a Hilbert space with norm $||\cdot|| = \sqrt{\langle \cdot,\cdot \rangle}$. Given $C \subset V$ a closed convex nonempty subset of V and $u_0 \in V$, there exists a unique element $\bar{u} \in C$ such that*

$$||u_0 - \bar{u}|| \le \inf_{v \in C} ||u_0 - v||.$$

We have $||u_0 - \bar{u}|| = d(u_0, C)$, that is, $\bar{u} \in C$ realizes the minimum of the distance between u_0 and C. We say that $\bar{u}$ is the projection of u_0 on C and we write

$$\bar{u} = proj_C\, u_0.$$

Moreover, $\bar{u}$ is characterized by the following property:

$$\begin{cases} \bar{u} \in C, \\ \langle u_0 - \bar{u}, v - \bar{u} \rangle \le 0 \quad \forall\, v \in C. \end{cases}$$

Because of the importance of this result, we give two proofs of independent interest. The first relies on Theorem 3.3.4 and is straightforward, but recall that we have used the weak topology to prove Theorem 3.3.4. The second is a direct one and completely elementary (it does not use the weak topology) and can be the starting point for developing a theory of Hilbert spaces at a more elementary level without using the weak topologies.

FIRST PROOF. First notice that it is equivalent to have

$$||u_0-\bar{u}||\le ||u_0-v|| \qquad \forall\, v\in C$$

or

$$||u_0-\bar{u}||^2\le ||u_0-v||^2 \qquad \forall\, v\in C.$$

So, we may say that our problem is equivalent to minimizing

$$f(v)=||u_0-v||^2+\delta_C(v)$$

over V. Clearly f is a convex function, as a sum of convex functions. It is strictly convex, because $||\cdot||^2$ is strictly convex and the sum of a convex and of a strictly convex function is still strictly convex. Since C is closed, δ_C is lower semicontinuous, and since $||\cdot||^2$ is continuous it is also lower semicontinuous, and f as a sum of two lower semicontinuous functions is still lower semicontinuous.

Finally $f(v)\ge ||u_0-v||^2$, which is clearly coercive, and so is f. So, f is strictly convex, lower semicontinuous, and coercive. It achieves its minimum at a unique point $\bar{u}\in C$.

SECOND PROOF. Let $(u_n)_{n\in\mathbf{N}}$ be a minimizing sequence, that is,

$$\begin{cases} u_n\in C,\\ ||u_0-u_n||^2\longrightarrow \inf\{||u_0-v||^2\ :\ v\in C\}=d(u_0,C)^2.\end{cases}$$

Then, use the parallelogram equality: given $n,m\in\mathbf{N}$,

$$2||u_0-u_n||^2+2||u_0-u_m||^2=||u_n-u_m||^2+4\left\|u_0-\frac{(u_n+u_m)}{2}\right\|^2.$$

Since C is convex, $(u_n+u_m)/2$ belongs to C and

$$\left\|u_0-\left(\frac{u_n+u_m}{2}\right)\right\|^2\ge d(u_0,C)^2.$$

Hence

$$||u_n-u_m||^2\le 2||u_0-u_n||^2+2||u_0-u_m||^2-4d(u_0,C)^2.$$

It follows that

$$\limsup_{n,m\longrightarrow+\infty}\ ||u_n-u_m||^2\le 0,$$

that is, the sequence $(u_n)_{n\in\mathbf{N}}$ is a Cauchy sequence. Since V is a Hilbert space, the sequence $(u_n)_{n\in\mathbf{N}}$ norm converges to some element $\bar{u}$ which still belongs to C, because C is closed. Moreover,

$$||u_0-\bar{u}||^2=\lim_n\ ||u_0-u_n||^2=d(u_0,C)^2.$$

Let us now prove the optimality condition for $\bar{u}$, that is,

$$\begin{cases} \bar{u}\in C,\\ \langle u_0-\bar{u},v-\bar{u}\rangle\le 0 \quad \forall v\in C.\end{cases}$$

This property says that $\bar{u}$ is characterized by the following geometrical property: for any $v\in C$, the angle between the two vectors $u_0-\bar{u}$ and $v-\bar{u}$ is greater than or equal to $\pi/2$.

We will later derive this property from general subdifferential calculus. At the moment, we give a direct elementary proof of it.

For any $v \in C$, by convexity of C, the line segment $[\bar{u}, v]$ still belongs to C and hence, for all $t \in [0,1]$,

$$w_t = tv + (1-t)\bar{u} \text{ belongs to } C.$$

By definition of $\bar{u}$

$$\begin{aligned} \|u_0 - \bar{u}\|^2 &\le \|u_0 - tv - (1-t)\bar{u}\|^2 \\ &\le \|(u_0 - \bar{u}) - t(v - \bar{u})\|^2. \end{aligned}$$

By developing this last expression, we obtain

$$2t\langle u_0 - \bar{u}, v - \bar{u}\rangle \le t^2 \|v - \bar{u}\|^2.$$

Divide by $t > 0$, and then let t go to zero to obtain

$$\langle u_0 - \bar{u}, v - \bar{u}\rangle \le 0.$$

Conversely, let us prove that if $\overline{u}$ satisfies the optimality condition above, then

$$\|u_0 - \overline{u}\| \le \inf_{v \in C} \|u_0 - v\|.$$

First notice that the optimality condition implies

$$\forall v \in C \quad \langle u_0 - v, v - \bar{u}\rangle \le 0. \tag{3.35}$$

Indeed,

$$\begin{aligned} \langle u_0 - v, v - \bar{u}\rangle &= \langle u_0 - \bar{u} + \bar{u} - v, v - \bar{u}\rangle \\ &= \langle u_0 - \bar{u}, v - \bar{u}\rangle - \|\bar{u} - v\|^2 \\ &\le 0. \end{aligned}$$

We then have

$$\begin{aligned} \|u_0 - \bar{u}\|^2 &= \langle u_0 - \bar{u}, u_0 - v + v - \bar{u}\rangle \\ &\le \langle u_0 - \bar{u}, u_0 - v\rangle \\ &= \langle u_0 - v + v - \bar{u}, u_0 - v\rangle \\ &= \|u_0 - v\|^2 + \langle u_0 - v, v - \bar{u}\rangle \\ &\le \|u_0 - v\|^2, \end{aligned}$$

where we have used the optimality condition in the first inequality and relation (3.35) in the last one. □

Corollary 3.3.2. *When $C = W$ is a closed subspace, then $\bar{u} = proj_W u_0$ is characterized by*

$$\begin{cases} \bar{u} \in W, \\ u_0 - \bar{u} \in W^{\perp}, \end{cases}$$

that is,

$$u_0 = (u_0 - \bar{u}) + \bar{u} \in W^{\perp} + W.$$

This is the orthogonal decomposition of $V = W \oplus W^{\perp}$ as the sum of two orthogonal subspaces. Moreover, the projection operator $proj : V \longrightarrow W$ is linear.

Proposition 3.3.2. *When V is a Hilbert space and C is a closed convex nonempty subset of V, the projection operator $V \longrightarrow C$ which associates to each $u \in V$ its projection $proj_C u$ on C is a contraction:*

$$\forall\, u, v \in V \quad ||proj_C u - proj_C v|| \leq ||u - v||.$$

PROOF. We have

$$\langle u - proj_C u, z - proj_C u\rangle \leq 0 \quad \forall\, z \in C,$$

$$\langle v - proj_C v, z - proj_C v\rangle \leq 0 \quad \forall\, z \in C.$$

Take $z = proj_C v$ in the first inequality and $z = proj_C u$ in the second one. Summing up, we obtain

$$\langle proj_C v - proj_C u, u - proj_C u - v + proj_C v\rangle \leq 0,$$

that is,

$$||proj_C v - proj_C u||^2 \leq \langle proj_C v - proj_C u, v - u\rangle.$$

By the Cauchy–Schwarz inequality, it follows that $||proj_C v - proj_C u|| \leq ||v - u||$. □

We end this section by remarking that the Hilbertian structure plays a fundamental role in the previous results for the best approximation. The existence of the best approximation $\bar{u}$ still holds true when $(V, ||\cdot||)$ is a reflexive Banach space. But when the space $(V, ||\cdot||)$ is no longer reflexive, even the existence of $\bar{u}$ may fail to be true, as shown by the following example.

Example 3.3.2. Take $V = \mathbf{C}([0,1];\mathbf{R})$ equipped with the sup norm

$$\forall v \in V \quad ||v||_\infty = \sup\{|v(t)| \,:\, t \in [0,1]\}.$$

Then $(V, ||\cdot||_\infty)$ is a Banach space which is not reflexive. Indeed, this is a consequence of the fact that the existence of a projection may fail to be true in this space: take

$$C = \left\{ v \in V \,:\, \int_0^{1/2} v(t)dt - \int_{1/2}^1 v(t)dt = 1 \right\}.$$

Clearly C is a closed convex nonempty subset of V (it is in fact a closed hyperplane). One can easily verify that $d(0, C) = 1$, but there is no element $v \in C$ such that $||v||_\infty = 1$. Indeed, if $v \in C$,

$$1 = \int_0^{1/2} v(t)dt - \int_{1/2}^1 v(t)dt \;\leq \frac{1}{2} \int_0^{1/2} |v(t)|dt + \frac{1}{2} \int_{1/2}^1 |v(t)|dt \leq ||v||_\infty.$$

Hence $d(0, C) = \inf\{||v||_\infty \,:\, v \in C\} \geq 1$, and it is not difficult to show that $d(0, C) = 1$. Suppose now that for some $v \in C$, $||v||_\infty = 1$. Since

$$\left| \int_0^{1/2} v(t)dt \right| \leq \frac{1}{2}, \qquad \left| \int_{1/2}^1 v(t)dt \right| \leq \frac{1}{2},$$

we necessarily have $\int_0^{1/2} v(t)dt = \frac{1}{2}$ and $\int_{1/2}^1 v(t)dt = -\frac{1}{2}$.

So $\int_0^{1/2}(1-v(t))dt=0$. Since $1-v(x)\geq 0$, this implies $v(t)\equiv 1$ on $[0,\frac{1}{2}]$. Similarly $v(t)\equiv -1$ on $[\frac{1}{2},1]$, a clear contradiction with the fact that v has to be continuous.

3.4 ▪ Ekeland's ε-variational principle

The so-called ε-variational principle was introduced by Ekeland in 1972 [206], [207], [208]. It is a general powerful tool in variational analysis and optimization which can be traced back to a maximality result for a partial ordering introduced by Bishop and Phelps in 1962 [100].

Ekeland's ε-variational principle asserts the existence of minimizing sequences of a particular kind. Not only do they approach the infimal value of the minimization problem, but they also simultaneously satisfy the first-order necessary conditions up to any desired approximation. In many instances, this makes this variational principle play a key role in the application of the direct method. Indeed, Ekeland's ε-variational principle has known considerable success with applications in a wide variety of topics in nonlinear analysis and optimization (critical point theory, geometry of Banach spaces, etc.).

In the last two decades, there has been increasing evidence that Ekeland's ε-variational principle has close connections with dissipative dynamical systems (dynamical systems with entropy). Indeed, solutions provided by the ε-variational principle can be seen as stable equilibria of such dynamics [138], [67]. In particular, the recent model in dynamical decision theory introduced by Attouch and Soubeyran [55] will serve as a guideline in the proof and interpretation of the results.

3.4.1 ▪ Ekeland's ε-variational principle and the direct method

Let us start with the following formulation of the Ekeland's ε-variational principle.

Theorem 3.4.1 (Ekeland). *Let (X,d) be a complete metric space and $f:X\longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function which is lower semicontinuous and bounded below ($\inf_X f>-\infty$). Then, for each $\varepsilon>0$, there exists some $x_\varepsilon\in X$ which satisfies the two following properties:*

$$\begin{cases} \text{(i)} & \inf_X f\leq f(x_\varepsilon)\leq \inf_X f+\varepsilon,\\ \text{(ii)} & f(x)\geq f(x_\varepsilon)-\varepsilon d(x,x_\varepsilon)\ \forall\, x\in X.\end{cases}$$

Let us first comment on this result and show some direct consequences of it. (Its proof is postponed to the next section.) Condition (ii) has a clear interpretation when $f:X\longrightarrow\mathbf{R}$ is Gâteaux differentiable on a Banach space $(X,||.||)$. It can be seen as a unilateral nonsmooth version of the condition $||Df(x_\varepsilon)||_*\leq\varepsilon$. Let us start with the definition of the Gâteaux differentiability property of f at x_ε. For any $\xi\in X$, with $||\xi||=1$ and any $t>0$,

$$f(x_\varepsilon+t\xi)=f(x_\varepsilon)+t\langle Df(x_\varepsilon),\xi\rangle+o(t).$$

By taking $x=x_\varepsilon+t\xi$ in (ii) and using the above equality, we obtain

$$f(x_\varepsilon)+t\langle Df(x_\varepsilon),\xi\rangle+o(t)\geq f(x_\varepsilon)-\varepsilon t.$$

Let us simplify, divide by $t>0$, and let $t\to 0^+$. We obtain

$$\langle Df(x_\varepsilon),\xi\rangle\geq-\varepsilon.$$

Changing ξ into $-\xi$ yields

$$|\langle Df(x_\varepsilon),\xi\rangle| \le \varepsilon.$$

This being true for any $\xi \in X$ with $||\xi|| \le 1$, we finally obtain $||Df(x_\varepsilon)||_* \le \varepsilon$.

We can summarize this result in the following corollary.

Corollary 3.4.1. *Let $(X, ||.||)$ be a Banach space and $f : X \longrightarrow \mathbf{R}$ a real-valued function which is lower semicontinuous, Gâteaux differentiable, and bounded below. Then, for each $\varepsilon > 0$, there exists some $x_\varepsilon \in X$ such that*

$$\begin{cases} \inf\limits_X f \le f(x_\varepsilon) \le \inf\limits_X f + \varepsilon, \\ ||Df(x_\varepsilon)||_* \le \varepsilon. \end{cases}$$

The above result asserts the existence of minimizing sequences $(x_n)_{n\in\mathbf{N}}$ of particular type: take $x_n = x_{\varepsilon_n}$ with $\varepsilon_n \to 0^+$; then

$$\begin{cases} f(x_n) \to \inf_X f \quad \text{as } n \to +\infty, \\ Df(x_n) \to 0 \quad \text{in } X^* \text{ as } n \to +\infty. \end{cases}$$

Application of the direct method when dealing with such particular minimizing sequences leads us naturally to introduce the so-called Palais–Smale compactness condition for a functional f.

Definition 3.4.1. *Let $(X, ||.||)$ be a Banach space. We say that a $\mathbf{C}^1$ function $f : X \longrightarrow \mathbf{R}$ satisfies the Palais–Smale condition if every sequence $(x_n)_{n\in\mathbf{N}}$ in X which satisfies*

$$\sup_n |f(x_n)| < +\infty \quad \textit{and} \quad Df(x_n) \to 0 \quad \textit{in } X^* \textit{ as } n \to +\infty$$

possesses a convergent subsequence (for the topology of the norm of X).

As an immediate consequence of Corollary 3.4.1, we obtain the next theorem.

Theorem 3.4.2. *Let $(X, ||.||)$ be a Banach space and $f : X \longrightarrow \mathbf{R}$ a $\mathbf{C}^1$ function which satisfies the Palais–Smale condition and which is bounded below. Then the infimum of f on X is achieved at some point $\bar{x} \in X$ and $\bar{x}$ is a critical point of f, i.e., $Df(\bar{x}) = 0$.*

Proof. Using Corollary 3.4.1 of the Ekeland's ε-variational principle, we have the existence of a sequence $(x_n)_{n\in\mathbf{N}}$ which satisfies

$$f(x_n) \to \inf_X f, \; Df(x_n) \to 0.$$

Since $\inf_X f \in_{\mathbf{R}}$, we have $\sup_n |f(x_n| < +\infty$, and the sequence $(x_n)_{n\in\mathbf{N}}$ satisfies the hypotheses of the Palais–Smale condition. Hence, one can extract a convergent subsequence $x_{n_k} \to \bar{x}$. By using the continuity properties of f and Df, one gets at the limit $f(\bar{x}) = \inf_X f$ and $Df(\bar{x}) = 0$. □

Judicious applications of this kind of result (based on the Palais–Smale compactness condition) provide existence results for critical points, not only local minima or maxima but also saddle points. One of the most celebrated of these results is the mountain pass theorem of Ambrosetti and Rabinowitz [15]. For further results in this direction, see [67] or [193].

Indeed, it turns out that when f is not necessarily smooth, condition (ii) is a convenient formulation of an ε-approximate optimality condition. The key is the following observation: property (ii) of Theorem 3.4.1 just expresses that x_ε is an exact solution of the perturbed minimization problem $(\mathscr{P}_\varepsilon)$:

$$\inf\{f(x)+\varepsilon d(x,x_\varepsilon): x\in X\}. \qquad (\mathscr{P}_\varepsilon)$$

In the particular and important case where f is convex and lower semicontinuous on a Banach space, one gets the following result.

Corollary 3.4.2. *Let $(X,||.||)$ be a Banach space and $f: X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function which is convex, lower semicontinuous, proper ($f \not\equiv +\infty$), and bounded below. Then, for each $\varepsilon > 0$ there exist $x_\varepsilon \in X$ and $x_\varepsilon^* \in X^*$ such that*

$$\begin{cases} \inf_X f \le f(x_\varepsilon) \le \inf_X f + \varepsilon, \\ x_\varepsilon^* \in \partial f(x_\varepsilon),\ ||x_\varepsilon||_* \le \varepsilon, \end{cases}$$

where $\partial f(x_\varepsilon)$ is the subdifferential of f at x_ε.

PROOF. We use standard tools from convex subdifferential calculus (see Chapter 9). Since x_ε minimizes the closed convex proper function $x \mapsto \varphi(x) := f(x)+\varepsilon||x-x_\varepsilon||$, we have $\partial\varphi(x_\varepsilon) \ni 0$. The norm being a continuous function in X, the additivity rule for the subdifferential calculus holds (Theorem 9.5.4) and we have

$$\partial f(x_\varepsilon)+\varepsilon\mathbf{B}(0,1)\ni 0.$$

Equivalently, there exists some $x_\varepsilon^* \in \partial f(x_\varepsilon)$ with $||x_\varepsilon^*||_* \le \varepsilon$. □

3.4.2 ▪ A dynamical approach and proof of Ekeland's ε-variational principle

Ekeland's ε-variational principle has a close connection with dissipative dynamical systems. This fact was recognized by Brezis and Browder [138]: "A general ordering principle"; Aubin and Ekeland [67]: "walking in complete metric spaces"; and Zeidler [363]: "The abstract entropy principle." More recently, the importance of this principle in the modelization of dynamical decision with bounded rationality was put to the fore by Attouch and Soubeyran [55]. This cognitive interpretation will serve as a guideline throughout this section.

The central concept in the dynamical approach to Ekeland's ε-variational principle is the following partial ordering relation.

Definition 3.4.2. *Let (X,d) be a metric space and $f: X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function which is proper ($f \not\equiv +\infty$). Let us introduce the following partial ordering on X:*

$$y \succeq_s x \iff f(y)+d(x,y)\le f(x).$$

We call it the marginal satisficing relation. We write

$$\begin{aligned} S(x) &= \{y\in X : y \succeq_s x\} \\ &= \{y\in X : f(y)+d(x,y)\le f(x)\} \end{aligned}$$

the set of elements of X which satisfy this ordering with respect to x.

Let us introduce some elements of decision theory that allow us to interpret this relation in a natural and intuitive way. Space X is the decision or performance space (the state space in physics). It is supposed that to each element $x \in X$ the agent is able to attribute a value or valence $f(x) \in \mathbf{R} \cup \{+\infty\}$ which measures the quality of the decision or performance x. (The value $+\infty$ allows us to take account of the constraints.) For example, when performing x, $f(x)$ measures how far the agent is from a given goal. In our context, $f(x)$ measures the dissatisfaction of the agent who, making $x \in X$, is faced with a problem which is not completely solved. Thus the agent is willing to reduce its dissatisfaction and make $f(x)$ as small as possible. The connection with the traditional formulation in decision sciences is obtained by taking $f(x) = \overline{g} - g(x)$, where g is a classical utility or gain function and $\overline{g}$ is a desirable level of resolution of the problem (for example, $\overline{g} = \sup_{x \in X} g(x)$). We choose this presentation to fit well with the classical formulation of variational principles in mathematics and physics and, in our situation, with the usual formulation of Ekeland's ε-variational principle.

The classical decision theory deals with perfectly rational agents who have immediate and free access to a global knowledge of their environment, and correspondingly minimize their value function f on X.

Modelization of decision processes in a complex real world requires us to introduce some further notions. Following Simon's [334] pioneering work in decision theory and bounded rationality, one needs to modelize the ability and difficulty of the agent to move and decide in a complex environment. A major difficulty for the agent is that it needs to explore its environment and get enough information to make further decisions. In this context, making decision becomes a dynamical process which at each step $k = 1, 2, \ldots$ is based on the following question: *Is it worthwhile for the agent to pass from a given state $x_k \in X$ (performance, decision, allocation) at time t_k into a further state $x_{k+1} \in X$ at time t_{k+1}?*

A key ingredient of the modelization of this balance between the advantage for the agent to pass from x to y and the possibility and difficulty of realizing it is the notion of cost to change. Following Attouch and Soubeyran [55], one introduces for any x and y in X, $c(x,y) \geq 0$, which is the cost to pass (change, move) from x to y. In our context, we assume that $c(x,y) \geq \theta d(x,y)$, where d is a metric on X and $\theta > 0$, is a unitary cost to move. This expresses that the cost to move is high for small displacements (by contrast with $c(x,y) = d(x,y)^2$, for example!). This metric d modelizes the difficulty for the agent to pass from a state x to a further state y. This is where the metric d in the cognitive interpretation of Ekeland's ε-variational principle appears! This is a rich concept which covers several aspects: there are costs to explore and get information (this comes from limited time and energy available for the agent), physical costs to move, and also costs with psychological and cognitive interpretation (dissimilarity costs, cost to quit a routine and enter into an other one, excitation and inhibition costs). From now on, we consider the particular situation $c(x,y) = d(x,y)$, which is enough for our purpose.

We now have the two terms of the balance: on one hand, the marginal gain $f(x) - f(y)$, and on the other hand, the cost to change $d(x,y)$. Precisely, the marginal satisficing relation $y \succeq_s x$ says that it is worthwhile for the agent to pass from x to y if the expected marginal gain ("reduction of dissatisfaction") $f(x) - f(y)$ is greater than or equal to the cost to pass from x to y:

$$f(x) - f(y) \geq d(x,y).$$

Let us examine some properties of the marginal satisficing relation $\succeq_s$.

Proposition 3.4.1. *The marginal satisficing relation $\succeq_s$ is a partial ordering relation on* $\operatorname{dom} f$*. An element $\bar{x} \in X$ is maximal with respect to this order iff for all $x \in X$, $x \neq \bar{x}$ one has*

$$f(\bar{x}) < f(x) + d(\bar{x}, x).$$

Proof. Clearly $\succeq_s$ is reflexive; we have $x \succeq_s x$ for all $x \in X$, because $d(x,x) = 0$.

Let us verify that $\succeq_s$ is antisymmetric. Suppose $y \succeq_s x$ and $x \succeq_s y$. We thus have

$$f(x) \geq f(y) + d(x,y) \text{ and } f(y) \geq f(x) + d(x,y).$$

Let us add these two inequalities and simplify the resulting expression. (At this point, note that it is important to consider x and y in $\operatorname{dom} f$.) One obtains $2d(x,y) \leq 0$, and hence $x = y$.

Let us now verify that $\succeq_s$ is transitive. Suppose that $z \succeq_s y$ and $y \succeq_s x$. We have

$$\begin{aligned} f(y) &\geq f(z) + d(y,z), \\ f(x) &\geq f(y) + d(x,y). \end{aligned}$$

By adding the two above inequalities, and using again that $x, y, z \in \operatorname{dom} f$, we obtain

$$\begin{aligned} f(x) &\geq f(z) + d(x,y) + d(y,z) \\ &\geq f(z) + d(x,z). \end{aligned}$$

In the above inequality we used the triangle inequality property satisfied by the metric d. Hence, $z \succeq_s x$ and $\succeq_s$ is transitive.

Let us now express that an element $\bar{x}$ of X is maximal with respect to the partial ordering relation $\succeq_s$. This means that for any $x \in X$, the following implication holds:

$$x \succeq_s \bar{x} \Longrightarrow x = \bar{x}, \quad \text{i.e.,}$$

$$\forall x \in X,\ x \neq \bar{x}, \quad f(\bar{x}) < f(x) + d(\bar{x}, x),$$

which completes the proof. □

Cognitive interpretation of maximal elements for $\succeq_s$. Let us show that maximal elements of the satisficing relation $\succeq_s$ can be interpreted as stable routines of the corresponding "worthwhile to move" (marginal satisficing) dynamical system. In our cognitive version, $\bar{x} \in X$ is said to be a stable routine if, starting from $\bar{x}$, the agent prefers to stay at $\bar{x}$ than to move from $\bar{x}$ to x for all x different from $\bar{x}$. Let us make this precise and consider, when starting from $\bar{x}$, the two following possibilities:

(a) if the agent chooses to stay at $\bar{x}$, his gain (dissatisfaction in our case) will be

$$f(\bar{x}) + d(\bar{x}, \bar{x}) = f(\bar{x});$$

(b) if the agent considers to move from $\bar{x}$ to x, his gain (dissatisfaction) is after moving (one adds two unsatisfactions, the cost to move $d(\bar{x}, x)$ and the dissatisfaction attached to x):

$$f(x) + d(\bar{x}, x).$$

Thus, the agent is willing to stay at $\bar{x}$ and rejects any move from $\bar{x}$ to x for all $x \neq \bar{x}$ iff

$$f(\bar{x}) < f(x) + d(\bar{x}, x),$$

which, by Proposition 3.4.1, just expresses that $\bar{x}$ is maximal for $\succeq_s$.

Therefore, Ekeland's ε-variational principle can be reformulated as an existence result of a maximal element for the partial ordering $\succeq_s$. Let us make this precise in the following statement.

Theorem 3.4.3. *Let us assume that (X,d) is a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is an extended real-valued function which is lower semicontinuous and bounded below. Then for any $x_0 \in \operatorname{dom} f$ there exists some $\bar{x} \in X$ which satisfies the two following properties:*

$$\begin{cases} \text{(i)} & \bar{x} \succeq_s x_0, \\ \text{(ii)} & \bar{x} \textit{ is maximal with respect to the partial ordering } \succeq_s. \end{cases}$$

Before proving Theorem 3.4.3, let us show how Ekeland's ε-variational principle can be derived from it: given $\varepsilon > 0$, take $x_0 \in \operatorname{dom} f$ such that

$$\inf_X f \le f(x_0) \le \inf_X f + \varepsilon.$$

Then, let us apply Theorem 3.4.3 with the metric εd and the corresponding satisficing relation

$$y \succeq_s x \iff f(y) + \varepsilon d(x,y) \le f(x).$$

Theorem 3.4.3 asserts the existence of $\bar{x}_\varepsilon$ such that $\bar{x}_\varepsilon \succeq_s x_0$ and $\bar{x}_\varepsilon$ maximal with respect to $\succeq_s$. The property $\bar{x}_\varepsilon \succeq_s x_0$ implies

$$\begin{aligned} f(\bar{x}_\varepsilon) &\le f(x_0) - \varepsilon d(\bar{x}_\varepsilon, x_0) \\ &\le f(x_0) \\ &\le \inf_X f + \varepsilon. \end{aligned}$$

On the other hand, by Proposition 3.4.1 and the maximality property of $\bar{x}_\varepsilon$, we have

$$\forall x \ne \bar{x}_\varepsilon \quad f(\bar{x}_\varepsilon) < f(x) + \varepsilon d(\bar{x}_\varepsilon, x).$$

Thus, $\bar{x}_\varepsilon$ satisfies the two desired properties (i) and (ii) of Theorem 3.4.1.

We are going to prove Theorem 3.4.3 (and hence Ekeland's ε-variational principle 3.4.1) by using the dynamical system, which is naturally associated to the marginal satisficing relation.

Definition 3.4.3. *A trajectory $(x_k)_{k\in\mathbf{N}}$ of the marginal satisficing dynamics is a sequence of elements x_k of X such that*

$$x_{k+1} \in S(x_k) \quad \forall\, k = 0,1,2,\ldots, \tag{S}$$

where S is the marginal satisficing relation. Equivalently, we have

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \cdots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \cdots,$$

that is,

$$f(x_{k+1}) + d(x_k, x_{k+1}) \le f(x_k) \quad \forall\, k = 0,1,2,\ldots.$$

Let us establish some general properties of the trajectories of the above dynamical system (S). We are mostly concerned with the asymptotic behavior as $k \to +\infty$ of these trajectories.

Proposition 3.4.2. *Let (X,d) be a metric space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function which is proper and bounded below. Take any trajectory $(x_k)_{k\in\mathbf{N}}$ of (S) starting from some $x_0 \in dom f$,*

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \cdots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \cdots.$$

Then, the following properties hold:

(i) *$(f(x_k))_{k\in\mathbf{N}}$ decreases with k, and $f(x_k)\to \inf_k f(x_k)\in\mathbf{R}$ when $k\to+\infty$.*

(ii) *The sequence $(x_k)_{k\in\mathbf{N}}$ satisfies $\sum_{k=0}^{+\infty} d(x_k,x_{k+1})<+\infty$. Hence, it is a Cauchy sequence in (X,d). When (X,d) is a complete metric space, the sequence $(x_k)_{k\in\mathbf{N}}$ converges in (X,d) to some $\bar{x}\in X$. Moreover, when f is lower semicontinuous, we have $\bar{x}\succeq_s x_k$ for all $k\in\mathbf{N}$.*

Proof. (i) For any $k\in\mathbf{N}$, by definition of $\preceq_s$

$$\begin{aligned} f(x_{k+1}) &\le f(x_{k+1})+d(x_k,x_{k+1})\\ &\le f(x_k). \end{aligned}$$

We have used $d(x_k,x_{k+1})\ge 0$, which expresses that changes are costly. Therefore, the sequence $(f(x_k))_{k\in\mathbf{N}}$ is decreasing. Since

$$-\infty<\inf_X f\le f(x_k)\le f(x_0)<+\infty,$$

we have $f(x_k)\downarrow \inf_k f(x_k)$, which is a finite real number.

(ii) Let us write the inequality $f(x_{k+1})+d(x_k,x_{k+1})\le f(x_k)$ for $k=0,1,\dots,n-1$,

$$\begin{aligned} &f(x_1)+d(x_0,x_1)\le f(x_0)\\ &\quad\vdots\\ &f(x_n)+d(x_{n-1},x_n)\le f(x_{n-1}). \end{aligned}$$

Then, we sum these inequalities and simplify the resulting expression. (Note that $f(x_k)\in\mathbf{R}$ for all $k\in\mathbf{N}$.) We obtain

$$f(x_n)+\sum_{k=0}^{n-1} d(x_k,x_{k+1})\le f(x_0).$$

Let us now use the minorization $f(x_n)\ge \inf_X f$ and the assumption $\inf_X f>-\infty$. We thus have

$$\sum_{k=0}^{n-1} d(x_k,x_{k+1})\le f(x_0)-\inf_X f<+\infty.$$

This being true for any $n\in\mathbf{N}$, we deduce

$$\sum_{k=0}^{+\infty} d(x_k,x_{k+1})\le f(x_0)-\inf_X f<+\infty.$$

Note that this holds true, just by assuming that (X,d) is a metric space. Then, by a classical argument, when (X,d) is a complete metric space, this implies the convergence of the sequence $(x_k)_{k\in\mathbf{N}}$ in (X,d). To see this, write the triangle inequality

$$\begin{aligned} d(x_n,x_{n+p}) &\leq \sum_{k=n}^{n+p-1} d(x_k,x_{k+1}) \\ &\leq \sum_{k=n}^{+\infty} d(x_k,x_{k+1}), \end{aligned}$$

which tends to zero as $n \to +\infty$. Hence $(x_k)_{k\in\mathbf{N}}$ is a Cauchy sequence in (X,d) which implies its convergence when (X,d) is a complete metric space. Let

$$x_k \to \bar{x} \text{ in } (X,d) \text{ as } k \to +\infty.$$

Let us prove that $\bar{x} \succeq_s x_n$ for all $n \in \mathbf{N}$. We have $x_k \succeq_s x_n$ for all $k \geq n$ (by transitivity of $\succeq_s$), i.e.,

$$f(x_k) + d(x_k,x_n) \leq f(x_n) \quad \forall\, k \geq n.$$

Let us fix $n \in \mathbf{N}$ and let $k \to +\infty$ in this inequality. Since $x_k \to \bar{x}$ in (X,d), by using the lower semicontinuity property of f (up to now we have not used it!), we obtain

$$f(\bar{x}) + d(\bar{x},x_n) \leq f(x_n),$$

that is, $\bar{x} \succeq_s x_n$. □

Our objective is to prove the existence of a trajectory $(x_k)_{k\in\mathbf{N}}$ of the marginal satisficing dynamics (S) which converges to a maximal element $\bar{x}$ for $\succeq_s$. So doing, we will have $\bar{x} \succeq_s x_k$ for all $k \in \mathbf{N}$ and hence $\bar{x} \succeq_s x_0$, which combined with the maximality of $\bar{x}$ is precisely the claim of Theorem 3.4.3.

In this perspective, to consider an arbitrary trajectory of (S) does not provide enough information: note that $x_k \equiv x_0$ for all $k \in \mathbf{N}$ is a trajectory of (S)! The dynamical system (S) modelizes a general rejection decision mechanism. We are now going to consider some trajectory of (S) which describes the decision process of a motivated agent. This means that at each step, the agent is willing to substantially improve his performance. This is a rich modelization subject involving some optimization aspects. In this perspective, the notion of aspiration index $m(x)$, which is defined in the next statement, plays an important role.

Lemma 3.4.1. *For any $x \in X$,*

$$\text{diam } S(x) \leq 2(f(x) - m(x)),$$

where

$$m(x) = \inf\{f(y) : y \in S(x)\} = \inf\{f(y) : y \succeq_s x\}$$

is called the aspiration index of the agent at x.

PROOF. The proof is an immediate consequence of the definition of $y \in S(x)$:

$$y \in S(x) \iff f(y) + d(x,y) \leq f(x).$$

Noticing that for $y \in S(x)$ we have $f(y) \geq m(x)$, we deduce

$$\forall y \in S(x) \qquad d(x,y) \leq f(x) - m(x).$$

As a consequence, for any $y, z \in S(x)$,

$$\begin{aligned} d(y,z) &\leq d(y,x) + d(x,z) \\ &\leq 2(f(x) - m(x)), \end{aligned}$$

and diam $S(x) \leq 2(f(x) - m(x))$. □

The aspiration index of the agent at x, say, $m(x)$, measures the gap between its present level of satisfaction at x and the maximum level of satisfaction that it can hope to obtain at a further step. Note that $m(x)$ is, in general, not known by the agent who is not able to explore all of $S(x)$. The cognitive model says that if the agent is motivated and is willing to explore enough at each step (and pay corresponding exploration costs!), then it knows a sufficiently good approximation of $m(x)$ and the process converges to a stable routine.

We are going to consider trajectories $(x_k)_{k \in \mathbf{N}}$ corresponding to a motivated agent who satisfies enough at each step. As an example, we consider that at each step k, the agent satisfies and fills a given fraction $\lambda \in]0,1[$ of the gap between $f(x_k)$ and $m(x_k)$: thus $x_{k+1} \succeq_s x_k$ and

$$f(x_{k+1}) \leq \lambda m(x_{k+1}) + (1-\lambda) f(x_k) = f(x_k) - \lambda[f(x_k) - m(x_k)].$$

We now have all the ingredients to state a dynamical, cognitive version and proof of Ekeland's ε-variational principle [55].

Theorem 3.4.4 (Attouch and Soubeyran). *Let (X,d) be a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is lower semicontinuous and bounded below.*

(a) *Then, for any $x_0 \in dom f$, there exists a trajectory $(x_k)_{k \in \mathbf{N}}$ of the marginal satisficing dynamical system (S),*

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \cdots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \cdots,$$

which converges in (X,d) to some $\bar{x} \in X$ which is a maximal element for the partial ordering $\succeq_s$.

(b) *Such trajectory can be obtained by satisficing enough at each step, for example, given some positive parameter $0 < \lambda < 1$, by taking at each step*

$$\begin{cases} x_{k+1} \succeq_s x_k \text{ and} \\ f(x_{k+1}) \leq f(x_k) - \lambda[f(x_k) - m(x_k)], \end{cases}$$

where $m(.)$ is the aspiration index : $m(x) = \inf\{f(y) : y \succeq_s x\}$.

Proof. Take a trajectory $(x_k)_{k \in \mathbf{N}}$ of (S) which satisfies enough at each step. One can always construct such a trajectory just by using the definition of $m(x)$ as an infimum. Then observe that the sequence $(S(x_k))_{k \in \mathbf{N}}$ is nested. Since $\succeq_s$ is transitive and $x_{k+1} \succeq_s x_k$, we have the following implication:

$$y \in S(x_{k+1}) \iff y \succeq_s x_{k+1} \Longrightarrow y \succeq_s x_k \iff y \in S(x_k),$$

i.e., $S(x_{k+1}) \subset S(x_k)$ for all $k \in \mathbf{N}$.

Let us prove that diam $S(x_k) \to 0$ as $k \to +\infty$. By using Lemma 3.4.1, it is enough to prove that $f(x_k) - m(x_k) \to 0$ as $k \to +\infty$.

Since $S(x_{k+1}) \subset S(x_k)$ we have

$$\begin{aligned} m(x_{k+1}) &= \inf\{f(y) : y \in S(x_{k+1})\} \\ &\geq \inf\{f(y) : y \in S(x_k)\} = m(x_k). \end{aligned}$$

We now use that this agent satisfies enough, i.e.,

$$f(x_{k+1}) \leq f(x_k) - \lambda[f(x_k) - m(x_k)],$$

and the inequality $m(x_{k+1}) \geq m(x_k)$ to obtain

$$\begin{aligned} f(x_{k+1}) - m(x_{k+1}) &\leq f(x_k) - \lambda[f(x_k) - m(x_k)] - m(x_k) \\ &\leq (1-\lambda)[f(x_k) - m(x_k)]. \end{aligned}$$

Hence

$$f(x_k) - m(x_k) \leq (1-\lambda)^k [f(x_0) - m(x_0)]$$

and

$$\text{diam } S(x_k) \leq 2(1-\lambda)^k [f(x_0) - m(x_0)].$$

Since $0 < \lambda < 1$ we have diam $S(x_k) \to 0$ as $k \to +\infty$. The sequence $(S(x_k))_{k \in \mathbf{N}}$ is a decreasing sequence of closed nonempty sets (closedness follows from the lower semicontinuity of f) whose diameter tends to zero. Since (X,d) is complete, we have, by a classical result, that $\bigcap_{k \in \mathbf{N}} S(x_k) = \{\bar{x}\}$ is nonvoid and is reduced to a single element $\bar{x} \in X$. For any $k \in \mathbf{N}$, we have x_k and $\bar{x}$, which belong to $S(x_k)$, hence $d(x_k, \bar{x}) \leq$ diam $S(x_k)$ which tends to zero. Thus, x_k converges to $\bar{x}$ in (X,d) as $k \to +\infty$.

The maximality of $\bar{x}$ with respect to $\succeq_s$ follows from the following observation: suppose that $y \succeq_s \bar{x}$. Since $\bar{x} \in S(x_k)$ for every $k \in \mathbf{N}$, we have $y \succeq_s x_k$ for all $k \in \mathbf{N}$, i.e., $y \in \bigcap_{k \in \mathbf{N}} S(x_k) = \{\bar{x}\}$. □

Indeed, when proving Theorem 3.4.3 and its dynamical version (Theorem 3.4.4), we have obtained a stronger version of Ekeland's variational principle, which is formulated below.

Theorem 3.4.5. *Let (X,d) be a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ a proper lower semicontinuous function which is bounded below. Let $\varepsilon > 0$ and $x_0 \in X$ be given such that*

$$f(x_0) \leq \inf_X f + \varepsilon,$$

and let $\lambda > 0$. Then there exists some $\bar{x}_{\varepsilon,\lambda} \in X$ such that

$$\begin{aligned} & f(\bar{x}_{\varepsilon,\lambda}) \leq f(x_0) \leq \inf_X f + \varepsilon; \\ & d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq \lambda; \\ & f(\bar{x}_{\varepsilon,\lambda}) < f(x) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x) \quad \forall\, x \neq \bar{x}_{\varepsilon,\lambda}. \end{aligned}$$

PROOF. Let us apply Theorem 3.4.3 with $\frac{\varepsilon}{\lambda} d$ instead of d. One obtains the existence of some $\bar{x}_{\varepsilon,\lambda} \in X$ which satisfies

$$\bar{x}_{\varepsilon,\lambda} \succeq_s x_0, \text{ i.e., } f(\bar{x}_{\varepsilon,\lambda}) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq f(x_0),$$

and $\bar{x}_{\varepsilon,\lambda}$ is maximal with respect to $\succeq_s$. We thus have $f(\bar{x}_{\varepsilon,\lambda}) \leq f(x_0)$ and

$$\begin{aligned}\inf_X f + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0) &\leq f(\bar{x}_{\varepsilon,\lambda}) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0)\\ &\leq f(x_0)\\ &\leq \inf_X f + \varepsilon,\end{aligned}$$

which implies $d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq \lambda$. The last property expresses that $\bar{x}_{\varepsilon,\lambda}$ is maximal with respect to $\succeq_s$. □