OPTIMISATION

IDRISS MAZARI-FOUQUER

Contents

Date: Last updated: September 18, 2025.

General introduction, pre-requisites

The goal of this course is to provide sound theoretical foundations for continuous optimisation. Optimisation problems arise everywhere, whether it be in medicine, in mechanics, in finance or in artificial intelligence. We will not be considering any specific application as we rather seek to provide a general overview of these problems and of the basic tools required to handle them.

We will consider optimisation problems that write

$$\min_{x \in X} f(x)$$

where X is a subset of a (possibly infinite dimensional) vector space and $f : X \to \mathbb{R}$ is the so-called "objective function". The class will proceed along the following lines:

(1) We will first review basic concepts in optimisation, the main goal being to provide standard tools to establish existence (and possibly uniqueness) of minimisers x^* of f. The key words are coercivity and convexity; a certain familiarity with differential calculus is required, as is a certain dexterity with linear algebra.

(2) The second part of the class will be dealing with unconstrained finite-dimensional optimisation, that is, when $X = \mathbb{R}^d$ for some $d \geq 1$. The main emphasis of this part will be on approximation procedures, the key concept being that of gradient descent. A reasonable familiarity with differential equations will be expected in the last section of this part.

(3) The third part of the class will be devoted to constrained optimisation problems, that is, when X is a strict subset of a vector space (still finite-dimensional at this stage). The key-concepts that the student should master at the end of this section are Lagrange multipliers and basic duality. We will also place a great emphasis on approximation methods, the two main algorithms being the projected gradient algorithm and the Uzawa algorithm.

(4) The final part of this class should be seen as an introduction on more advanced topics in optimisation in the infinite dimensional setting: calculus of variations, optimal control and back-propagation, the latter being of central importance in the study of Neural Networks.

Parts of these lectures are inspired by classes I gave at other points, parts are inspired by the works of my predecessors (P. Cardaliaguet, Y. Viossat).The class should be self-contained and, as usual, there is no need to use other references: the lectures and exercise sessions should provide enough material. Nevertheless, some "standard" references are the following:

References

[1] Nicolas Boumal. *An introduction to optimization on smooth manifolds.* Cambridge: Cambridge University Press, 2023.

[2] Nicolas Boumal, Dmitriy Drusvyatskiy, and Quentin Rebjock. Gradient descent can converge to any isolated saddle point.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[4] Guillaume Carlier. *Classical and modern optimization.* Adv. Textb. Math. Hackensack, NJ: World Scientific, 2022.

[5] Guillaume Garrigos and Robert M. Gower. Handbook of Convergence Theorems for (Stochastic) Gradient Methods. Preprint, arXiv:2301.11235 [math.OC] (2023), 2023.

Part 1. **A review of basic concepts**

1. Existence of optimisers, optimality conditions

Throughout this entire chapter, unless stated otherwise, we consider a fixed, $\mathscr{C}^2$ function $f : \mathbb{R}^d \to \mathbb{R}$.

1.0.1. *First definitions.* The goal of this section is to fix the terminology, as well as some notations.

Definition 1.1. *A point $x^* \in \mathbb{R}^d$ is called:*

(1) *A* global minimiser of f *if*
$$\forall x \in \mathbb{R}^d, f(x^*) \leq f(x).$$
(2) *A* local minimiser of f *if*
$$\exists \varepsilon > 0, \forall x \in \mathbb{R}^d, \|x - x^*\| \leq \varepsilon \Rightarrow f(x^*) \leq f(x).$$

In the remainder of this document, we will adopt the following notational conventions:

(1) $M_d(\mathbb{R})$ denotes the set of $d \times d$ matrices, $M_{p,q}(\mathbb{R})$ denotes the set of $p \times q$ matrices.
(2) $S_d(\mathbb{R})$ denotes the set of symmetric matrices in $M_d(\mathbb{R})$. The transpose of a matrix M is written M^T.
(3) $S_d^+(\mathbb{R})$, resp. $S_d^{++}(\mathbb{R})$, resp. $S_d^-(\mathbb{R})$, resp. $S_d^{--}(\mathbb{R})$ denotes the set of symmetric positive, resp. definite positive, resp. negative, resp. symmetric negative, matrices.

1.1. **The optimisation problem under consideration.** The goal of this class is to study the optimisation problem

$$\inf_{x \in \mathbb{R}^d} f(x). \tag{1.1}$$

A first question is whether or not a solution x^* actually exists, in which case we are allowed to write

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

To this end, let us recall the definition of *coercivity*:

Definition 1.2. *We say that a continuous function f is* coercive *if for any $M \in \mathbb{R}$ the sub-level set $\{x \in \mathbb{R}^d : f(x) \leq M\}$ is bounded.*

The main point of coercivity is the following proposition:

Proposition 1.1. *Assume f is coercive. Then f has a global minimiser x^*. With a slight abuse of terminology, we will say that x^* solves (1.1), and dub it a* minimiser *of f in $\mathbb{R}^d$.*

Proof of Proposition 1.1. Let $x_0 \in \mathbb{R}^d$ be arbitrary and consider the set

$$E_0 := \{x \in \mathbb{R}^d : f(x) \leq f(x_0)\}.$$

By assumption, E_0 is a compact set and it is clear that if x^* solves

$$\inf_{x \in E_0} f(x) \tag{1.2}$$

then it solves (1.1). However, the existence of a solution to (3.2) follows from the Weierstraßtheorem: every continuous function on a compact set reaches its extrema on this set. □

1.2. **Optimality conditions.** Now, consider (1.1), and assume that x^* is a solution. Naturally, one would like to have either an explicit or a good enough numerical approximation of the minimiser x^*. Unless we are quite lucky and an easy comparison argument provides an explicit value, this is a hopeless endeavour. The best we can do is to rely on *optimality conditions*. Optimality conditions allow to reduce the search of a minimiser to the resolution of a non-linear system of equations.

There are two optimality conditions. The first-order optimality condition reads

$$\nabla f(x^*) = 0 \tag{1.3}$$

while the second-order optimality condition writes

$$\nabla^2 f(x^*) \in S_d^+(\mathbb{R}).$$

Assume for simplicity that (1.3) has a finite number of solutions $x_1, \dots, x_N$, which have tractable expressions. This still does not provide any conclusion, and we need to compute the Hessian of f at each x_k, $k = 1, \dots, N$. There are several possibilities, summarised in the following proposition:

For the sake of future references, let us single out the following definition:

Definition 1.3. *Let $f \in \mathscr{C}^1(\mathbb{R}^d; \mathbb{R})$. A point $x^* \in \mathbb{R}^d$ is called a* critical point *of f if*

$$\nabla f(x^*) = 0.$$

Proposition 1.2. *Assume f is $\mathscr{C}^2$ and let x^* be a critical point of f.*

1. *If $\nabla^2 f(x^*) \in S_d^{++}(\mathbb{R})$, then x^* is a strict local minimiser of f.*
2. *If $\nabla^2 f(x^*) \in S_d^{--}(\mathbb{R})$, then x^* is a strict local maximiser of f.*
3. *If $\nabla^2 f(x^*)$ has at least one negative and one positive eigenvalue, x^* is a* saddle *point: there exist two orthogonal directions $\vec{e}_1, \vec{e}_2$ such that $t^* = 0$ is a local minimiser of $t \mapsto f(x+te_1)$, and a local maximiser of $t \mapsto f(x+te_2)$.*
4. *If $\nabla^2 f(x^*) \in S_d^+(\mathbb{R})$, but not in $S_d^{++}(\mathbb{R})$, then we cannot conclude and further analysis is required.*

In the first two cases, we say that x^* is a non-degenerate critical point.

Proof of Proposition 1.2. We only prove the first and third points. We first assume that

$$\nabla^2 f(x^*) \in S_d^{++}(\mathbb{R}). \tag{1.4}$$

There are two ways to prove that x^* is a local minimiser, both of which naturally rely on Taylor expansions, and on the following consequence of (1.4) (see Exercise 1.1): there exists a constant $\underline{\lambda} > 0$ such that, for any $z \in \mathbb{R}^d$,

$$\langle \nabla^2 f(x^*) z, z \rangle \geq \underline{\lambda} \|z\|^2. \tag{1.5}$$

(1) First approach: a proof by contradiction Argue by contradiction and assume that there exists a sequence $\{x_k\}_{k \in \mathbb{N}}$ such that

$$\forall k \in \mathbb{N}, f(x_k) \leq f(x^*).$$

From the mean-value formula, we know that for any $k \in \mathbb{N}$ there exists $\xi_k \in [x_k; x^*] = \{(1-t)x_k + tx^*, t \in [0;1]\}$ such that

$$f(x_k) = f(x^*) + \frac{1}{2} \langle \nabla^2 f(\xi_k)(x_k - x^*), x_k - x^* \rangle.$$

In particular, setting for any $k \in \mathbb{N}$ $z_k := \frac{x_k - x^*}{\|x_k - x^*\|}$,

$$\langle \nabla^2 f(\xi_k) z_k, z_k \rangle \leq 0.$$

As for any $k \in \mathbb{N}$ we have $\|z_k\| = 1$ we can (up to taking a subsequence) assume that $\{z_k\}_{k\in\mathbb{N}}$ converges to some z_∞ , $\|z_\infty\| = 1$. Since $x_k \underset{k\to\infty}{\to} x^*$, it follows that $\xi_k \underset{k\to\infty}{\to} x^*$. Passing to the limit in the previous inequality we obtain

$$\langle \nabla^2 f(x^*) z_\infty, z_\infty \rangle \leq 0,$$

in contradiction with (1.5).

(2) Second approach: continuity of the Hessian From (1.5) and the fact that $\nabla^2 f$ is continuous it is possible to show the following fact (see Exercise 1.1): there exists $\underline{\lambda}' > 0$ and $\varepsilon > 0$ such that

$$\forall x \in \mathbb{B}(x^*; \varepsilon)\,, \forall z \in \mathbb{R}^d\,, \langle \nabla^2 f(x) z, z \rangle \geq \underline{\lambda}' \|z\|^2.$$

We can conclude as before: fix such an $\varepsilon > 0$. Then, for any $x \in \mathbb{B}(x^*; \varepsilon)$, there exists $\xi \in \mathbb{B}(x^*; \varepsilon)$ such that

$$f(x) - f(x^*) = \frac{1}{2} \langle \nabla^2 f(\xi)(x - x^*), (x - x^*) \rangle \geq \frac{\underline{\lambda}'}{2} \|x - x^*\|^2$$

and the conclusion follows.

□

The exercises of this chapter (in particular Exercise 1.3) contain several examples of optimisation problems that can be solved by hand. Such examples are usually limited to dimension 2 or 3, unless the problem has a very specific structure.

2. Convexity, strict convexity

The second crucial notion is that of convexity, which has two main interests, as we shall explain:

(1) The first one is when studying global properties of a functions: in that case, assuming convexity helps in either proving uniqueness of the optimiser (in the strongly convex case) or at least to have some geometric structure on the set of minimisers.
(2) The second one is more local in nature: consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that admits a non-degenerate local minimum $x^* \in \mathbb{R}^d$. Then, locally around x^*, by Taylor expansions, f can be suitably approximated by a (strongly) convex functions. Thus, we can hope that any global study of convex functions can translate to local studies around non-degenerate local minima (this is of particular importance when dealing with gradient descents).

2.1. Various definitions of convexity and basic properties.

Definition 2.1. *A set $K \subset \mathbb{R}^n$ is said to be convex if for all x and y in K, $tx+(1-t)y \in K$ for all t in $[0,1]$ (for any two points in K, the segment that unites them is in K).*

Definition 2.2. *Let $K \subset \mathbb{R}^n$ be a convex set and $f : K \to \mathbb{R}$ be a function.*

(1) *f is **convex** on K if*

$$\forall x\,, y \in K,\ t \in (0,1)\,, f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

(2) *f is **strictly convex** on K if*

$$\forall x \neq y \in K,\ t \in (0,1)\,, f(tx + (1-t)y) < tf(x) + (1-t)f(y).$$

(3) *f is **strongly convex** on K if there exists $\alpha > 0$ such that*

$$\forall x, y \in K,\ t \in [0,1]\,, f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\alpha}{2}t(1-t)\|x-y\|^2.$$

(4) *f is said to be **concave** if $-f$ is convex (and similar definitions for strictly or strongly concave).*

Of course, these definitions can be a tad annoying to work with. When f is more regular (say, $\mathscr{C}^1$ or $\mathscr{C}^2$), equivalent characterisations are available.

Proposition 2.1 (Equivalent characterisation of convexity in the $\mathscr{C}^1$ regime)**.** *Let $f \in \mathscr{C}^1(\mathbb{R}^d; \mathbb{R})$. Then the following propositions are equivalent:*

(1) *f is convex.*
(2) *(A convex function is above its tangent hyperplane) For any $x\,, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{2.1}$$

(3) *(The gradient of a convex function is monotone) For any $x\,, y \in \mathbb{R}^d$,*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0. \tag{2.2}$$

Proof of Proposition 2.1. Assume that f is convex and let $x\,, y \in \mathbb{R}^d$. Thus, for any $t \in [0;1]$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

or, alternatively,

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

As f is $\mathscr{C}^1$, passing to the limit $t \to 0$ gives

$$\langle \nabla f(x), y - x \rangle \leq f(y) - f(x),$$

which is exactly (2.1). Now, assuming (2.1), we obtain

$$\langle \nabla f(x), y - x \rangle \leq f(y) - f(x) \leq -\langle \nabla f(y), x - y \rangle$$

or, equivalently,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0.$$

Now, assuming (2.2), let us show (2.1). Let $x\,, y \in \mathbb{R}^d$. Then, by the Taylor formula,

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f\left((1-t)x + ty\right), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f\left((1-t)x + ty\right) - \nabla f(x), y - x \rangle dt + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

However, setting $x_t := (1-t)x + ty$, we have

$$x_t - x = t(y - x).$$

In particular,

$$\langle \nabla f(x_t) - \nabla f(x), y - x \rangle \geq 0$$

whence

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$$

so that (2.1) is satisfied. Finally, assume (2.1). Let us show that f is convex. Let $x\,, y \in \mathbb{R}^d$ and $t \in [0;1]$. Consider the map

$$g : t \mapsto f\left((1-t)x + ty\right) - (1-t)f(x) - tf(y).$$

Then $g(0) = g(1) = 0$ and, retaining the notation $x_t := (1-t)x + ty$,

$$g'(t) = \langle \nabla f(x_t), y - x \rangle - (f(y) - f(x)).$$

Observe that for any $t_0\,, t_1 \in [0;1]$ there holds

$$\begin{aligned} (t_0 - t_1)(g'(t_0) - g'(t_1)) &= \langle \nabla f(x_{t_0}) - \nabla f(x_{t_1}), (t_0 - t_1)(y - x) \rangle \\ &= \langle \nabla f(x_{t_0}) - \nabla f(x_{t_1}), x_{t_0} - x_{t_1} \rangle \\ &\geq 0 \end{aligned}$$

whence g' is non-decreasing. By the Rolle theorem, there exists $s \in (0;1)$ such that $g'(s) = 0$. Consequently, g is non-increasing on $(0;s)$ and non-decreasing on $(s;1)$, and is thus maximal at either $t = 0$ or $t = 1$, thereby concluding the proof of convexity of f. □

We leave as an exercise the following proposition:

Proposition 2.2 (Equivalent characterisation of strong convexity in the $\mathscr{C}^1$ regime)**.** *Let $f \in \mathscr{C}^1(\mathbb{R}^d; \mathbb{R})$. Then the following propositions are equivalent:*

(1) *f is α-strongly convex convex.*
(2) *For any $x\,, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{2.3}$$

(3) *For any $x\,, y \in \mathbb{R}^d$, $x \neq y$*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2. \tag{2.4}$$

To conclude these reminders, we recall some characterisation of convex $\mathscr{C}^2$ functions:

Proposition 2.3. *Let $f \in \mathscr{C}^2(\mathbb{R}^d;\mathbb{R})$. Then the following properties are equivalent:*

(1) *f is convex.*
(2) *For any $x \in \mathbb{R}^d\,, \nabla^2 f(x) \in S_d^+(\mathbb{R})$.*

There is also a nice characterisation of α-strongly convex functions:

Proposition 2.4. *Let $f \in \mathscr{C}^2(\mathbb{R}^d;\mathbb{R})$ and $\alpha > 0$. Then the following properties are equivalent:*

(1) *f is α-strongly convex.*
(2) *For any $x \in \mathbb{R}^d$, the lowest eigenvalue $\lambda_1(\nabla^2 f(x))$ of the Hessian of f satisfies*

$$\lambda_1\left(\nabla^2 f(x)\right) \geq \alpha.$$

Unfortunately, there is no such nice characterisation of strict convexity, but merely an implication (we also refer to Exercise 1.5):

Proposition 2.5. *Let $f \in \mathscr{C}^2(\mathbb{R}^d;\mathbb{R})$. If*

$$\forall x \in \mathbb{R}^d\,, \nabla^2 f(x) \in S_d^{++}(\mathbb{R})$$

then f is strictly convex.

Similarly, we leave the proofs of these propositions as exercises.

2.2. **Convex function and minimisation.** The main result of this section is the following:

Theorem 2.1. *Let f be a convex function, and assume that the problem*

$$\min_{x\in\mathbb{R}^d} f(x) \tag{2.5}$$

has a solution x^*. Then:

(1) *If f is strictly convex, (2.5) has a unique solution.*
(2) *In general, the set of minimisers*

$$X := \{x \in \mathbb{R}^d :\ f(x) = f(x^*)\}$$

is a convex set.

Proof of Theorem 2.1. We begin with the fact that the set of minimisers X is convex: let $x\,, y$ be such that $f(x) = f(y) = f(x^*)$. Then, by convexity of f,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) = f(x^*),$$

whence $(1-t)x + ty \in X$. Now, suppose that f is strictly convex and assume by contradiction that f has two distinct minimisers $x\,, y$. By strict convexity, for any $t \in (0;1)$,

$$f((1-t)x + ty) < (1-t)f(x) + tf(y) = f(x^*),$$

in contradiction with the minimality of x^*. □

2.3. **Several (useful) inequalities related to convex functions.** To conclude the first part, we will give and prove several inequalities related to convex functions– these might seem quite abstruse at first, but they will be coming in handy in later parts of the class, so that this specific section of the lecture notes should be taken as a reference point for later purposes. Furthermore, the proofs allow to get familiar with usual tricks when dealing with convex functions.

Proposition 2.6. *Let f be a convex function and assume that ∇f is μ-Lipschitz continuous. Then*

$$\forall x\,, y \in \mathbb{R}^d\,, f(x) - f(y) - \langle \nabla f(y), y - x\rangle \geq \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2. \tag{2.6}$$

In particular, if f admits a minimiser x^*, it follows that

$$\forall x \in \mathbb{R}^d\,, f(x) - f(x^*) \geq \frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{2.7}$$

This is sometimes referred to as "co-coercivity of the gradient".

Proof of Proposition 2.6. The key is to introduce an auxiliary point z. Fix $x\,, y \in \mathbb{R}^d$. Then for any $z \in \mathbb{R}^d$

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\leq \langle \nabla f(x), x - z\rangle \\ &\quad + f(z) - f(y) \\ &\quad \text{by convexity} \\ &\leq \langle \nabla f(x), x - z\rangle + \langle \nabla f(y), z - y\rangle + \frac{\mu}{2}\|z - y\|^2 \\ &\quad \text{by } \mu\text{-Lipschitzianity of the gradient.} \end{aligned}$$

We now minimise the right-hand side with respect to z. As $\varphi : z \mapsto \langle z, \nabla f(y) - \nabla f(x)\rangle + \frac{\mu}{2}\|y - z\|^2$ is a strictly convex function of z, if z^* is a critical point of φ, then it is a global minimiser of φ. As

$$\nabla\varphi(z^*) = 0 \Leftrightarrow z^* = y + \frac{1}{\mu}\left(\nabla f(x) - \nabla f(y)\right)$$

we obtain

$$f(x) - f(y) \leq \langle \nabla f(x), x\rangle - \langle \nabla f(y), y\rangle + \varphi(z^*).$$

Expanding, we deduce

$$\begin{aligned} f(x) - f(y) &\leq \langle \nabla f(x), x - y\rangle - \frac{1}{\mu}\langle \nabla f(x), \nabla f(x) - \nabla f(y)\rangle \\ &\quad + \frac{1}{\mu}\langle \nabla f(y), \nabla f(x) - \nabla f(y)\rangle + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \langle \nabla f(x), x - y\rangle - \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2, \end{aligned}$$

which is the desired inequality. □

The final inequality, the proof of which is to be found in Exercise 1.9, is the Polyak-Lojasiewicz inequality:

Proposition 2.7. *Let $f \in \mathscr{C}^1(\mathbb{R}^d; \mathbb{R})$ be an α-strongly convex function. Then*

$$\forall x \in \mathbb{R}^d\,, f(x) - \inf_{\mathbb{R}^d} f \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2.$$

Exercises for Part 1

Exercise 1.1. *Let $A \in S_d(\mathbb{R})$.*

(1) *Letting $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_d(A)$ be the eigenvalues of A, show that*
$$\lambda_1(A) = \inf_{\|z\|^2=1} \langle Az, z\rangle.$$
(2) *Show that for any two $A, B \in S_d(\mathbb{R})$ there holds*
$$|\lambda_1(A) - \lambda_1(B)| \leq \|A - B\|_{\mathrm{op}}$$
where $\|\cdot\|_{\mathrm{op}}$ stands for the standard operator norm on the set of matrices.

Exercise 1.2. *Let $A \in S_d(\mathbb{R})$ and $b \in \mathbb{R}^d$. We consider*
$$f : x \mapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle.$$
(1) *Show that f is coercive if, and only if $A \in S_d^{++}(\mathbb{R})$.*
(2) *Show that f is convex if, and only if $A \in S_d^{+}(\mathbb{R})$.*
(3) *Show that f is strictly convex if, and only if $A \in S_d^{++}(\mathbb{R})$.*

Exercise 1.3. *Classify the critical points (local minimisers, local maximisers, saddle points, indeterminate critical points) of the following functions:*

(1) $f_1 : (x, y) \mapsto (x - y)^2 + (x + y)^3$,
(2) $f_2 : (x, y) \mapsto x^2 - 2y^2 + 3xy$,
(3) $f_3(x, y) \mapsto x^4 + y^3 - 3y - 2$.

Exercise 1.4 (Distance between two sets)**.** *Let A and B be two closed, nonempty subsets of $\mathbb{R}^d$.*

(1) *Show that if A is compact, then the problem*
$$\min_{a\in A,\, b\in B} \|a - b\|$$
has a solution (at least one).
(2) *Show with a counter-example that this problem need not have a solution if neither A nor B is assumed compact, even if A and B are convex.*

Exercise 1.5. *Give an example of a strictly convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that the equation*
$$\nabla^2\varphi(x) = 0$$
has infinitely many solutions.

Exercise 1.6 (Carathéodory theorem)**.** *Let $\Omega \subset \mathbb{R}^d$. We call the convex hull of Ω the smallest convex set containing Ω. We denote it by $C(\Omega)$.*

(1) *Show that*
$$C(\Omega) = \left\{ \sum_{i=0}^{N} t_i x_i \,,\, N \in \mathbb{N} \,,\, \{t_i\}_{i=0,\ldots,N} \in [0;1]^{N+1} \,,\, \sum_{i=0}^{N} t_i = 1 \,,\, \{x_i\}_{i=0,\ldots,N} \in \Omega^{N+1} \right\}.$$
(2) *We now want to show the Carathéodory theorem: for any $x \in C(\Omega)$, there exist $t_0, \ldots, t_d \in [0; 1], x_0, \ldots, x_d \in \Omega$ such that*
$$\sum_{i=0}^{d} t_i = 1 \,, x = \sum_{i=0}^{d} t_i x_i.$$

(a) *Using an example, show why one needs at least* $(d+1)$ *points.*
(b) *Prove the Carathéodoory theorem; you can argue by descending induction, starting (for instance) from a point* $x \in C(\Omega)$ *that writes* $x = \sum_{i=0}^{d+1} t_i x_i$*, and showing that one of the vectors* $x_i's$ *can be expressed using the others.*
(c) *Deduce from the Carathéodory theorem that if* Ω *is compact, then so is* $C(\Omega)$*.*

Exercise 1.7 (Extreme points I: projection on closed convex sets)**.** *Let* $K \subset \mathbb{R}^d$ *be a closed convex set. Show that there exists a unique* $z \in K$*, denoted by* Π_K *and dubbed the orthogonal projection of* x *on* K*, such that*

$$\|x - \Pi_K(x)\| = \min_{z \in K} \|x - z\|$$

and that

$$\forall y \in K\,, \langle x - \Pi_K(x), y - \Pi_k(x)\rangle \leq 0.$$

Show that Π_K *is* 1*-Lipschitz.*

Exercise 1.8 (Extreme points II: The Krein-Milman theorem)**.**
(1) *Give an example of a convex set* $K \subset \mathbb{R}^d$ *that has no extreme points.*
(2) *We assume that* K *is compact. Prove that* K *has extreme points.*
(3) *We now want to prove the (finite-dimensional) Krein-Milman theorem: any* $x \in K$ *is a convex combination of extreme points of* K*.*
   (a) *Let* $x \in \partial K$*. Show that there exists a hyperplane (called the supporting hyperplane)* $H = \{\varphi = 0\}$ *where* $\varphi \in (\mathbb{R}^d)'\,, \varphi \neq 0$ *(the dual of* $\mathbb{R}^d$*) such that* $x \in H$ *and* $\varphi(K) \subset (-\infty; 0]$ *(hint: think of the projection* Π_K*).*
   (b) *Let* $x \in K$*. Show that if* $x \in H$ *for some supporting hyperplane of* K *(i.e. associated with some* $y \in K$*) then* x *is an extreme point of* K *if, and only if,* x *is an extreme point of* $H \cap K$*.*
   (c) *Show the Krein-Milman theorem proceeding by induction on the dimension.*

Exercise 1.9 (Polyak-Lojasciewicz Inequality)**.** *Let* $f : \mathbb{R} \to \mathbb{R}$ *be an* α*- strongly convex function and let* x^* *be a minimiser of* f*. First, prove that*

$$\forall x \in \mathbb{R}^d\,, \|x - x^*\|^2 \leq \frac{2}{\alpha}(f(x) - f(x^*)).$$

Second, show the following inequality:

$$\forall x \in \mathbb{R}^d\,, f(x) - f(x^*) \leq \frac{1}{2\alpha} \|\nabla f(x)\|^2\,.$$

Finally, deduce that

$$\forall x \in \mathbb{R}^d\,, \|x - x^*\| \leq \frac{1}{\alpha}\|\nabla f(x)\|.$$

Part 2. Unconstrained optimisation: approximation and gradient descent

3. An overview of gradient descent: basic principle & study of convex functions

3.1. **Goal of the part.** In this part of the class we will study a first approximation method for optimisation problems. In other words, consider once again a problem of the form

$$\min_{x\in\mathbb{R}^d} f(x) \tag{3.1}$$

and assume for the sake of simplicity that a solution x^* exists. As we already mentioned, the first exploitable information on x^* we have is that it is a critical point of f:

$$\nabla f(x^*) = 0.$$

The goal of the gradient descent is to **find critical points of** f through an iterative method, that is, a method which can be written as

$$\begin{cases} \text{Start from an initial guess } x_0, \\ \text{Supposing } x_0, \dots, x_k \text{ are built, set } x_{k+1} = x_k + G_k(x_k) \end{cases}$$

for some function G_k, the definition of which might depend on the previous iterates $x_0, \dots, x_k$. Most of the time, this will not be the case, and the iteration map G_k will not depend on the index k. The goal is to obtain algorithms that produce sequences that converge at a "good enough" rate — most of the time, we will be satisfied with *linear convergence*, in the following sense:

Definition 3.1. *Let $\{x_k\}_{k\in\mathbb{N}} \in (\mathbb{R}^d)^{\mathbb{N}}$ and $x^* \in \mathbb{R}^d$. We say that $\{x_k\}_{k\in\mathbb{N}}$ converges linearly, at rate $\alpha \in [0;1)$, to x^*, if there exists a constant C such that*

$$\forall k \ \in \mathbb{N}\,, \|x_k - x^*\| \leq C\alpha^k.$$

3.2. **Definition of the gradient descent and basic properties.**

3.2.1. *First considerations.* The gradient descent is a *local* algorithm that essentially relies on a Taylor expansion of the function f: assume that you are starting from an initial guess $x_0 \in \mathbb{R}^d$, and you want to solve (1.1). We look for a point that is close enough to x_0, say at distance at most d_0, and such that $f(x_1) < f(x_0)$ (if that is possible). In that case, a natural idea is to replace f by its first-order Taylor approximation

$$f(x_0 + z) = f(x_0) + \langle \nabla f(x_0), z\rangle + \underset{z\to 0}{o}(\|z\|)$$

so that, at first order, we are solving the minimisation problem

$$\min_{\|x-x_0\|\leq d_0} \langle \nabla f(x_0), x - x_0\rangle. \tag{3.2}$$

At this stage, two things may happen:

(1) Either the gradient vanishes ($\nabla f(x_0) = 0$), in which case we stop, as we are satisfied with what we already have. Now, if we wanted to go further in the analysis, we should note that two other possibilities arise: either f is convex, in which case this implies that x_0 is a global minimiser of the function f, or f is not convex and we would need to do something different

to investigate the local optimality of x_0. This will very often not be the case.

(2) Either the gradient does not vanish, so that the pseudo-optimisation problem (3.2) has a unique solution

$$x_1 = -\frac{d_0}{\|\nabla f(x_0)\|}\nabla f(x_0).$$

Now, the question remains of choosing the parameter d_0. Of course, if we already know that the gradient is small enough in norm, it makes no sense to look for a point that would be far away, and this naturally leads to choosing d_0 as $d_0 = \tau\|\nabla f(x_0)\|$ for some $\tau > 0$.

Overall, we define the sequence of iterates of the gradient descents as follows:

$$\begin{cases} x_0 \in \mathbb{R}^d, \\ \forall k \in \mathbb{R}^d, x_{k+1} = x_k - \tau\nabla f(x_k). \end{cases}$$

The main questions under consideration from now on are:

(1) The **convergence** of the generated sequence $\{x_k\}_{k\in\mathbb{N}}$.
(2) The **convergence of the sequence of values** $\{f(x_k)\}_{k\in\mathbb{N}}$.
(3) The **convergence of the gradient of the objective function** $\{\nabla f(x_k)\}_{k\in\mathbb{N}}$.

Of course, the convergence of $\{x_k\}_{k\in\mathbb{N}}$ implies the convergence of the values and of the gradient; the convergence of the values, on the other hand, does not imply the convergence of the sequence itself. It is also important to note that, in general, the presentation of gradient descent assumes, from the get-go, some strong convexity of f, which gives a positive answer to all the questions above. On the other hand, it is extremely important, both in practice and in theory, to distinguish these different steps and this is what we will do. At any rate, here is a simple result:

Proposition 3.1. *Assume that $f \in \mathscr{C}^1$ and that the gradient descent with fixed step size τ converges in the sense that $\{x_k\}_{k\in\mathbb{N}}$ converges to some x^*. Then $\nabla f(x^*) = 0$.*

Proof of Proposition 3.1. If the sequence converges then, passing to the limit in $x_{k+1} = x_k - \nabla f(x_k)$ yields $\nabla f(x^*) = 0$. □

Of course the next question is, if we assume that $\{x_k\}_{k\in\mathbb{N}}$ converges to some x^*, is it true that x^* is, in fact, a minimiser of f? The answer is no in general. Consider for instance the function

$$f : x \mapsto \frac{x^3}{3},$$

a fixed $1 > \tau > 0$ and an initialisation $x_0 = 1$. The sequence of iterates of the gradient descent is given by

$$\forall k \in \mathbb{N}, x_{k+1} = x_k(1 - \tau x_k).$$

Now, if $\tau \in (0; 1)$, a simple reasoning by induction shows that the sequence $\{x_k\}_{k\in\mathbb{N}}$ is positive and decreasing; in particular it is converging so that by Proposition 3.1 it converges to 0, which is not a minimiser of f. Of course, one might argue that this is cheating, as the function f is not coercive. Nevertheless, it is easy to adapt this example: simply modify f on $(-\infty; -1]$ to have a globally smooth, coercive function.

3.2.2. *Do the step-size and the regularity matter?* In this first paragraph, we investigate in a formal manner the constraints we should put on the step size and on the function f to obtain a converging sequence, where the parameters should be chosen uniformly with respect to the initial condition. We begin with the regularity of the function. Let us consider the case of a $\mathscr{C}^1$, but not $\mathscr{C}^{1,1}$ function, for instance, in two variables

$$f : (x, y) \mapsto \frac{2}{3}\left(x^2 + 2y^2\right)^{\frac{3}{4}}.$$

It is fairly easy to show that the function f is $\mathscr{C}^1$, but not $\mathscr{C}^{1,1}$ at 0 (this is left as an exercice): simply observe that

$$|f(x, y)| \leq C\|(x, y)\|^{\frac{3}{2}}.$$

Furthermore, 0 is the unique minimiser of f. For a given parameter $\tau > 0$, the sequence of iterates is given explicitly by

$$\begin{cases} x_{k+1} = x_k \left(1 - \frac{\tau}{(x_k^2+2y_k^2)^{\frac{1}{4}}}\right) \\ y_{k+1} = y_k \left(1 - \frac{2\tau}{(x_k^2+2y_k^2)^{\frac{1}{4}}}\right). \end{cases}$$

Observe that, at a formal level, if the sequence converges, then it must converge to 0. Thus, we "should" be able to write that

$$(x_{k+1}, y_{k+1}) \sim \frac{\tau}{(x_k^2 + 2y_k^2)^{\frac{1}{4}}} (-x_k, -2y_k).$$

Defining

$$z_k := x_k^2 + 2y_k^2$$

we deduce that (asymptotically)

$$z_{k+1} \geq C z_k^{\frac{1}{2}}, C = \tau^2.$$

Now, let us assume that this inequality is, in fact, satisfied for all $k \in \mathbb{N}$. This would give the lower bound

$$z_{k+1} \geq C C^{\frac{1}{2}} z_{k-1}^{\frac{1^2}{2^2}} \geq \cdots \geq C^{\sum_{i=0}^{k} \left(\frac{1}{2}\right)^i} z_0^{2^{-k}},$$

and cannot converge to 0.

We continue with an investigation of the step size; here the computations are much easier, as it suffices to consider, in the one-dimensional case, the function

$$f : x \mapsto \frac{\mu}{2} x^2.$$

Then, for any initialisation x_0 and any fixed step size $\tau > 0$, the sequence of iterates is given by

$$x_{k+1} = x_k(1 - \mu\tau) = x_0(1 - \mu\tau)^k.$$

Thus, the method converges if, and only if, $0 < \tau < \frac{1}{\mu}$. As μ quantifies the steepness of f', or, put otherwise, the average variation of the gradient, we fairly easily understand that the wilder the gradient of a function, the smaller the step size needs to be.

3.2.3. *The gradient descent is a descent method.* In this section, we consider a function $f \in \mathscr{C}^1(\mathbb{R}^d)$ with a μ-Lipschitz gradient in the sense that

$$\forall x\,, y \in \mathbb{R}^d\,, \|\nabla f(x) - \nabla f(y)\| \leq \mu \|x - y\|. \tag{3.3}$$

We do not make any assumption on the coercivity of f, or on the existence of a minimiser. Our first result is the following:

Theorem 3.1. *For any $x_0 \in \mathbb{R}^d$, for any $\tau > 0$, the sequence generated by the gradient descent initialised at x_0 with step size τ satisfies*

$$\forall k \in \mathbb{N}\,, f(x_{k+1}) - f(x_k) \leq \tau\left(\tau\mu - 1\right)\|\nabla f(x_k)\|^2.$$

In particular, if $\tau \in \left(0; \frac{1}{\mu}\right)$ then the sequence $\{f(x_k)\}_{k\in\mathbb{N}}$ is strictly decreasing unless it is stationary. Finally, for any $\tau \in \left(0; \frac{1}{2\mu}\right)$ there holds

$$\forall k \in \mathbb{N}\,, f(x_{k+1}) - f(x_k) \leq -\frac{\tau}{2}\|\nabla f(x_k)\|^2.$$

Proof of Theorem 3.1. It suffices to write that for any $k \in \mathbb{N}$ there holds

$$f(x_{k+1}) = f(x_k - \tau\nabla f(x_k)).$$

From the mean-value theorem, there exists $\xi \in \mathbb{B}(x_k; \|x_{k+1} - x_k\|)$ such that

$$f(x_{k+1}) = f(x_k) + \langle \nabla f(\xi), -\tau\nabla f(x_k)\rangle.$$

This rewrites

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= -\langle \nabla f(\xi) - \nabla f(x_k), \tau\nabla f(x_k)\rangle - \tau\|\nabla f(x_k)\|^2 \\ &\leq \tau\|\nabla f(\xi) - \nabla f(x_k)\| \cdot \|\nabla f(x_k)\| - \tau\|\nabla f(x_k)\|^2 \\ &\leq \tau\mu\|x_{k+1} - x_k\| \cdot \|\nabla f(x_k)\| - \tau\|\nabla f(x_k)\|^2 \\ &= \tau\left(\tau\mu - 1\right)\|\nabla f(x_k)\|^2. \end{aligned}$$

The conclusion follows. □

We highlight once again that we did not require any information other than the regularity of ∇f. In the next section, we will illustrate several nice properties of gradient descent when the function f to be optimised is convex.

3.2.4. *Convergence of the gradient descent II: convex functions.* We now make one stronger assumption on the function f. Namely, we assume that f still satisfies (3.3) for some constant $\mu > 0$ and that f is convex.

Theorem 3.2. *Assume that f is convex and satisfies (3.3) for some $\mu > 0$. Finally, assume that f has a minimiser x^*. For any $\tau \in \left(0; \frac{1}{2\mu}\right)$, for any initialisation x_0, the gradient descent with fixed step size τ, initialised at x_0, satisfies*

$$\forall k \in \mathbb{N}\,, f(x_{k+1}) - f(x^*) \leq \frac{\|x_k - x^*\|^2}{2\tau(k+1)}.$$

Proof of Theorem 3.2. Recall that from Theorem 3.1 we have

$$\forall k \in \mathbb{N}\,, f(x_{k+1}) - f(x_k) \leq -\frac{\tau}{2}\|\nabla f(x_k)\|^2.$$

However, by convexity of f,

$$f(x_k) \leq f(x^*) + \langle \nabla f(x_k), x_k - x^*\rangle.$$

Consequently,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{\tau}{2}\|\nabla f(x_k)\|^2 \\ &\leq f(x^*) + \langle \nabla f(x_k), x_k - x^* \rangle - \frac{\tau}{2}\|\nabla f(x_k)\|^2 \\ &= f(x^*) + \frac{2}{\tau}\left(\frac{\tau}{2}\langle \nabla f(x_k), x_k - x^* \rangle - \frac{\tau^2}{4}\|\nabla f(x_k)\|^2\right) \\ &= f(x^*) - \frac{2}{\tau}\left(\|\frac{\tau}{2}\nabla f(x_k) - \frac{1}{2}(x_k - x^*)\|^2 - \frac{1}{4}\|x_k - x^*\|^2\right) \\ &= f(x^*) - \frac{2}{\tau}\left(\frac{1}{4}\|x_{k+1} - x^*\|^2 - \frac{1}{4}\|x_k - x^*\|^2\right). \end{aligned}$$

We thus deduce that

$$k\left(f(x_k) - f(x^*)\right) \leq \sum_{i=1}^{k}(f(x_i) - f(x^*)) \leq \frac{1}{2\tau}\|x_0 - x^*\|^2.$$

The conclusion follows.

□

3.2.5. *Convergence of the gradient descent III: quadratic functions.* We saw in the previous paragraph that, in the case of convex functions, we could get a convergence rate (algebraic, as it turns out) for the gradient descent. The goal of this section is to provide a finer convergence rate in the special case of quadratic functions.

Definition 3.2. *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is quadratic if there exists $A \in M_d(\mathbb{R})$ and $b \in \mathbb{R}^d$ such that*

$$\forall x \in \mathbb{R}^d, f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle.$$

When f is quadratic, we say that f is represented by (A, b).

A straightforward computation shows that

$$\nabla f(x) = \frac{A + A^T}{2}x - b.$$

In particular, when A is symmetric,

$$\nabla f(x) = Ax - b$$

and x^* is a critical point of f if, and only if, x^* is a solution to $Ax^* = b$.

Theorem 3.3. *Let $A \in S_d^{++}(\mathbb{R})$, $b \in \mathbb{R}^d$ and f be the quadratic function represented by (A, b). Letting $0 < \lambda_1 \leq \lambda_d(A)$ be the eigenvalues of A, for any $\tau \in \left(0; \frac{2}{\lambda_d(A)}\right)$, for any $x_0 \in \mathbb{R}^d$, the gradient descent initialised at x_0 with fixed step size $\tau > 0$ converges linearly to the unique solution of $Ax^* = b$ and, more specifically,*

$$\forall k \in \mathbb{N}, \|x_k - x^*\| \leq \alpha(\tau)^k \|x_0 - x^*\|$$

with $\alpha(\tau) = \max_{i=1,\dots,d}|1 - \tau\lambda_i(A)|$. Finally,

$$\min_{\tau \in \left(0; \frac{2}{\lambda_d(A)}\right)} \alpha(\tau) = \alpha\left(\frac{1}{\lambda_1(A) + \lambda_d(A)}\right) = \frac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A) + 1} \text{ with } \mathrm{cond}(A) = \frac{\lambda_d(A)}{\lambda_1(A)}.$$

Proof of Theorem 3.3. Observe that as $A \in S_d^{++}(\mathbb{R})$ all the eigenvalues of A are positive. Furthermore, A induces a bijection, whence x^* is uniquely defined. Additionally, as A is symmetric, $\nabla f(x) = Ax - b$. Now we explicitly obtain, for any $k \in \mathbb{N}$,

$$x_{k+1} = x_k - \tau A x_k + \tau b$$

so that, defining $y_k := x_k - x^*$,

$$\forall k \in \mathbb{N}\,, y_{k+1} = y_k - \tau A x_k + \tau b = y_k - \tau A(x_k - x^*) = (\mathrm{Id} - \tau A) y_k.$$

The matrix $\mathrm{Id} - \tau A$ has eigenvalues $1 - \tau\lambda_d(A) \leq \cdots \leq 1 - \tau\lambda_1(A)$We deduce that

$$\forall k \in \mathbb{N}\,, \|y_{k+1}\| = \|(\mathrm{Id} - \tau A) y_k\| \leq \|\mathrm{Id} - \tau A\|_{\mathrm{op}} \cdot \|y_k\|.$$

Here, we used the operator norm on $\mathrm{Id} - \tau A$. By a straightforward iteration argument we deduce that

$$\forall k \in \mathbb{N}\,, \|y_k\| \leq \|\mathrm{Id} - \tau A\|_{\mathrm{op}}^k \, \|y_0\|.$$

However,

$$\|\mathrm{Id} - \tau A\|_{\mathrm{op}} = \max_{i=1,\dots,d} |1 - \tau\lambda_i(A)|.$$

We refer to Exercise 1.1. In particular, if $\tau > 0$ is chosen so that

$$\alpha(\tau) := \max_{i=1,\dots,d} |1 - \tau\lambda_i(A)| < 1 \tag{3.4}$$

we obtain

$$\forall k \in \mathbb{N}\,,\ \|x_k - x^*\| \leq \alpha(\tau)^k \|x_0 - x^*\|.$$

It remains to pick $\tau > 0$ so that $\alpha(\tau) < 1$. However

$$\alpha(\tau) < 1 \Leftrightarrow -1 < 1 - \tau\lambda_d(A) \leq 1 - \tau\lambda_1(A) < 1$$

which rewrites, in a compact form, as

$$\tau < \frac{2}{\lambda_d(A)}.$$

The conclusion follows. Finally, it is an easy exercise to see that α is minimised at τ^* such that

$$|1 - \tau^*\lambda_1(A)| = |1 - \tau^*\lambda_d(A)|.$$

Solving this equation explicitly in τ^* yields

$$\tau^* = \frac{2}{\lambda_1(A) + \lambda_d(A)}\,,\ \text{whence } \alpha(\tau^*).$$

□

Let us observe that the convergence rate of gradient descent is quantified by the conditioning number of the matrix A: if $\mathrm{cond}(A) \approx 1$ then the method converges extremely quickly if τ^* is chosen properly, while, if $\mathrm{cond}(A) \gg 1$ (which means that A, as a linear map, dilates much more in certain directions than in others), the method will *a priori* converge extremely slowly. It is important to have basic reflexes regarding the conditioning number of matrix. We refer to Exercise 2.3.

3.2.6. *Convergence of the gradient descent IV: the case of strongly convex functions.* The purpose of this section is to generalise the results of the previous paragraph to the case of strongly convex functions:

Theorem 3.4. *Let f be a α-strongly convex, coercive, $\mathscr{C}^1$ function with a μ-Lipschitz gradient. Let x^* be the minimiser of f. Then, for any $x_0 \in \mathbb{R}^d$, for any $\tau \in \left(0; \frac{1}{2\mu}\right)$, the gradient descent initialised at x_0 with fixed step size τ converges linearly to x^* and, more precisely, we have*

$$\forall k \in \mathbb{N}\,, \|x_k - x^*\| \leq (1 - \alpha\tau)^{\frac{k}{2}} \|x_0 - x^*\|.$$

Proof of Theorem 3.4. We observe that, setting $y_k := x_k - x^*$, we have

$$\forall k \in \mathbb{N}\,, y_{k+1} = y_k - \tau \left(\nabla f(x_k) - \nabla f(x^*)\right).$$

Taking the squared norm on each side of this identity yields

$$\|y_{k+1}\|^2 = \|y_k\|^2 + \tau^2 \|\nabla f(x_k)\|^2 - 2\tau \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle$$

Now observe that

$$\frac{\alpha}{2}\|x_k - x^*\|^2 + \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

so that

$$\|y_{k+1}\|^2 \leq (1 - \alpha\tau)\|y_k\|^2 + 2\tau(f(x^*) - f(x_k)) + \tau^2 \|\nabla f(x_k)\|^2$$

From (2.7)

$$\tau^2 \|\nabla f(x_k)\|^2 \leq 2\tau^2 \mu (f(x_k) - f(x^*)).$$

Consequently

$$\|y_{k+1}\|^2 \leq (1 - \alpha\tau)\|y_k\|^2 + 2\tau \left(1 - \tau\mu\right)\left(f(x^*) - f(x_k)\right) \leq (1 - \alpha\tau)\|y_k\|^2.$$

□

4. Beyond convexity

One might then wonder what happens when the function f to be optimised is no longer convex. We will be going over two phenomena: the first one is that, even when f is not convex, the gradient descent converges locally around non-degenerate minimisers–this is absolutely expected. The second phenomenon is different in nature, and amounts to investigating whether we can guarantee that the gradient descent at least converges. It is the case, provided f only has isolated critical points. This relies on the Zoutendijk theorem.

4.1. **Convergence around non-degenerate local minimisers.** The first theorem that we give, the proof of which is given in Exercise 2.4, is the following:

Theorem 4.1. *Let $f \in \mathscr{C}^2(\mathbb{R}^d;\mathbb{R})$ have a μ-Lipschitz gradient. Let x^* be a non-degenerate local minimiser of f in the sense that $\nabla f(x^*) = 0\,, \nabla^2 f(x^*) \in S_d^{++}(\mathbb{R})$. There exists $\varepsilon > 0$ such that, for any $\tau > 0$ small enough, for any $x_0 \in \mathbb{B}(x^*;\varepsilon)$, the gradient descent initialised at x_0 with step size τ converges linearly to x^*.*

This is unsurprising: locally around x^*, one can write

$$f(x) \approx f(x^*) + \frac{1}{2}\langle \nabla^2 f(x^*)(x-x^*), x-x^*\rangle$$

so the situation "should" resemble the quadratic case. The only difficulty is in controlling the error term in the approximation above. The proof is carried out in Exercise 2.4.

4.2. **Convergence to isolated critical points.** A much more interesting extension of the analysis we carried out above is the case of non-convex functions. Observe that, in general, one can not obtain any convergence result, *per* the counter examples we already encountered. Nevertheless, if the function f is coercive and has only isolated critical points, it is possible to reach the following conclusion:

Theorem 4.2. *Let $f \in \mathscr{C}^1(\mathbb{R}^d;\mathbb{R})$ be coercive and have a μ-Lipschitz gradient. Assume that f only has isolated critical points. Then for any $\tau > 0$ small enough, for any $x_0 \in \mathbb{R}^d$, the sequence of iterates generated by the gradient descent initialised at x_0 with step-size τ converges, and its limit is a critical point of f.*

Proof of Theorem 4.2. The proof is quite lengthy, but elementary. We split it into several parts for the convenience of the reader:

(1) The sequence $\{x_k\}_{k\in\mathbb{N}}$ remains bounded Recall that if $\tau \in \left(0; \frac{1}{2\mu}\right)$ there holds

$$f(x_{k+1}) - f(x_k) \leq \frac{-\tau}{2}\|\nabla f(x_k)\|^2.$$

In particular, $\{x_k\}_{k\in\mathbb{N}} \subset \{f \leq f(x_0)\} = K_0$, which is, by assumption, a compact set. The conclusion follows.

(2) There holds $\|x_{k+1} - x_k\| \underset{k\to\infty}{\to} 0$ Similarly, summing the previous estimates, we obtain

$$\frac{\tau}{2}\sum_{k=0}^{\infty}\|\nabla f(x_k)\|^2 \leq f(x_0) - f(x^*),$$

which implies

$$\|x_{k+1} - x_k\|^2 = \|\nabla f(x_k)\|^2 \underset{k\to\infty}{\to} 0. \tag{4.1}$$

(3) The sequence $\{x_k\}_{k\in\mathbb{N}}$ has a unique closure point Of course, this property would suffice to conclude, as any sequence living in a compact set converges if, and only if, it has a unique closure point. However, observe that from (4.1) any closure point of $\{x_k\}_{k\in\mathbb{N}}$ is a critical point. As critical points are isolated, and as K_0 is compact, there are finitely many critical points $y_0, \dots, y_N$ in K_0. By continuity of the gradient, we can fix $\varepsilon, \delta > 0$ such that

$$\forall i \neq j\,, \overline{\mathbb{B}}(y_i;\varepsilon) \cap \overline{\mathbb{B}}(y_j;\varepsilon) = \emptyset\,, \inf_{\overline{K_0 \setminus \cup_i \mathbb{B}(y_i;e)}} \|\nabla f\| \geq \delta > 0.$$

I particular, if $\{x_k\}_{k\in\mathbb{N}}$ had two distinct closure points, say y_0 and y_1, it would follow from (4.1) that there exists a subsequence $\{x_{n_k}\}_{k\in\mathbb{N}} \subset \overline{K_0 \setminus \cup_i \mathbb{B}(y_i;e)}$, which would contradict that any closure point is critical. Consequently, $\{x_k\}_{k\in\mathbb{N}}$ has a unique critical point, and thus converges.

□

Exercises for Part 2

Exercise 2.1. (1) *We let* $A = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ *and* f *be represented by* (A, b). *Can the gradient descent initialised at a given* $x_0 \in \mathbb{R}^d$ *with fixed step size* $\tau > 0$ *converge?*

(2) *Assume that* A *is symmetric and that for any* $b \in \mathbb{R}^d$, *for any* $x_0 \in \mathbb{R}^d$ *there exists* $\tau > 0$ *such that the gradient descent generated at* x_0 *with step size* $\tau > 0$ *converges. Show that* $A \in S_d^{++}(\mathbb{R})$.

Exercise 2.2. *We let* $A \in S_d(\mathbb{R})$ *be matrix with (at least) two eigenvalues of opposite signs. We let* $b = 0$. *Show that for any* $\tau > 0$ *the set* $\{x_0 \in \mathbb{R}^d$: *the gradient descent initialised at* x_0 *with fixed step size* τ *converges*$\}$ *has measure zero.*

Exercise 2.3. *[Some basic properties of the conditioning number]*

(1) *Show that, for any symmetric positive definite matrix* M, $\mathrm{cond}(M) \geq 1$.
(2) *Show that for any symmetric definite positive matrix* $\mathrm{cond}(M) = \|M\|_{\mathrm{op}} \cdot \|M^{-1}\|_{\mathrm{op}}$. *We use this expression to define the conditioning number of any invertible matrix* $M \in Gl_d(\mathbb{R})$.
(3) *Show that for any* $M \in Gl_d(\mathbb{R})$ $\mathrm{cond}(M) \geq 1$ *and that, for any orthogonal matrix* P, $\mathrm{cond}(PM) = \mathrm{cond}(M)$.
(4) *For any* $M \in Gl_d(\mathbb{R})$ *show that* $\|M\|_{\mathrm{op}} = \|M^T\|_{\mathrm{op}}$.
(5) *Let* $M \in Gl_d(\mathbb{R})$ *be such that* $\mathrm{Cond}(M) = 1$. *Show that there exists* $x \in \mathbb{R}^*$ *such that* xM *is an orthogonal matrix.*

Exercise 2.4. *Prove Theorem 4.1.*

Exercise 2.5. *Let* $f \in \mathscr{C}^1(\mathbb{R}^d; \mathbb{R})$ *be bounded from below, satisfy the Polyak-Lojasiewicz condition with constant* α*:*

$$\forall x \in \mathbb{R}^d, f(x) - \inf_{\mathbb{R}^d} f \leq \frac{1}{2\alpha} \|\nabla f(x)\|^2.$$

Assume that ∇f *is* μ*-Lipschitz. For any* $\tau \in \left(0; \frac{1}{\mu}\right)$ *any* $x_0 \in \mathbb{R}^n$, *let* $\{x_k\}_{k \in \mathbb{N}}$ *be the sequence generated by the gradient descent initialised at* x_0 *with fixed step size* τ. *Show that*

$$\forall k \in \mathbb{N}, f(x_{k+1}) - \inf f \leq (1 - \tau\alpha)^{k+1} (f(x_0) - \inf f).$$

Exercise 2.6. *The goal of this exercise is to show the convergence of the line-search gradient descent for quadratic functions.*

(1) *Preliminary: Kantorovich inequality Let* $A \in S_d^{++}(\mathbb{R})$ *with eigenvalues* $0 < \lambda_1 \leq \cdots \leq \lambda_d$. *Show that*

$$\forall x \in \mathbb{R}^d \setminus \{0\}, \|x\|^4 \leq \langle Ax, x\rangle \cdot \langle A^{-1}x, x\rangle \leq \frac{\|x\|^4}{4} \cdot \frac{(\lambda_1 + \lambda_d)^2}{\lambda_1 \lambda_d}.$$

(2) *Let* $A \in S_d^{++}(\mathbb{R})$ *and* $b \in \mathbb{R}^d$. *Let* $x \in \mathbb{R}^d$. *Solve the optimisation problem*[1]

$$\min_{\tau > 0} f(x - \tau \nabla f(x)).$$

[1] In particular, show existence and uniqueness of the optimiser

(3) *We now consider the sequence generated by the line search algorithm. Using the explicit expression of the step size obtained at the previous question and defining, for any* $k \in \mathbb{N}$, $y_k := A(x_k - x^*)$, *show that*

$$\forall k \in \mathbb{N}\,, \langle y_{k+1}, x_{k+1} - x^* \rangle = \langle y_k, x_k - x^* \rangle \cdot \left(1 - \frac{\|y_k\|^4}{\langle Ay_k, y_k \rangle \langle A^{-1} y_k, y_k \rangle}\right).$$

(4) *Conclude the proof.*