Chapter 2
Basics in optimization

Nelly Pustelnik

Goals of the chapter.

- Review the framework (Hilbert spaces, linear operators, adjoint, gradient, Hessian, Taylor expansions).
- Introduce the classes of functions used in optimization: smooth (β-smooth) functions and functions in $\Gamma_0(\mathcal{H})$ (proper, lower semicontinuous, convex).
- Characterize minimizers: existence, uniqueness, first- and second-order optimality conditions.
- Examples of minimization problems.

Contents

1 Framework and reminders **2**

1.1 Hilbert spaces . . . 2
1.2 Optimization problems . . . 2
1.3 Gradient and Hessian matrix . . . 3
1.4 Norms in finite dimension . . . 3
1.5 Linear operators, norm and adjoint . . . 4
1.6 Approximation by Taylor expansion . . . 6

2 Classes of functions **7**

2.1 Continuously differentiable functions . . . 8
2.2 Domain, epigraph, lower semicontinuity . . . 8
2.3 Convexity . . . 9
2.4 Strict and strong convexity . . . 11
2.5 Characterizations of convexity . . . 11
2.6 Weak convexity . . . 12

3 Minimizers and optimality conditions **13**

3.1 Local and global minimizers, stationary points . . . 13
3.2 How to recognize a minimum? . . . 13
3.3 The convex case . . . 15
3.4 Examples . . . 16

4 Iterative descent methods **16**

4.1 Descent directions . . . 17
4.2 What do we expect from an optimization method? . . . 18

5 Concrete examples **18**

5.1 Inverse problems in imaging . . . 18
5.2 Image segmentation . . . 21
5.3 Supervised learning . . . 21

6 Exercises **22**

We think that convex optimization is an important enough topic that everyone who uses computational mathematics should know at least a little bit about it. In our opinion, convex optimization is a natural next topic after advanced linear algebra and linear programming.
– Stephen Boyd and Lieven Vandenberghe –

Whatever people do, at some point they get a craving to organize things in a best possible way. This intention, converted in a mathematical form, turns out to be an optimization problem of certain type.
– Yurii Nesterov –

1 Framework and reminders

1.1 Hilbert spaces

Definition 1.1 – (Hilbert space)

A (real) Hilbert space $\mathcal{H}$ is a complete real vector space endowed with an inner product $\langle \cdot \,|\, \cdot \rangle$. The associated norm is

$$(\forall x \in \mathcal{H}) \qquad \|x\| = \sqrt{\langle x \,|\, x \rangle}.$$

The fundamental special case is $\mathcal{H} = \mathbb{R}^N$ (Euclidean space of dimension N), endowed with the inner product $\langle x \,|\, y \rangle = x^\top y = \sum_{i=1}^N x_i y_i$. **Throughout this course, we work in finite dimension.**

1.2 Optimization problems

Let $f : \mathcal{H} \to \mathbb{R} : x \mapsto f(x)$. We consider *unconstrained* optimization problems, that is, problems of the form

$$\min_{x \in \mathcal{H}} f(x)$$

or, stated differently, problems that consist in finding

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x).$$

The function f is called the *objective function* (or criterion). We will see in Section 2 that this formulation also makes it possible to handle constraints, by allowing f to take the value $+\infty$ (indicator function).

Example 1.1 (Quadratic functions)**.** A quadratic function is a function of the form

$$f : \mathbb{R}^N \to \mathbb{R}, \qquad f(x) = \tfrac{1}{2} x^\top A x - b^\top x = \tfrac{1}{2} \langle x \,|\, Ax \rangle - \langle b \,|\, x \rangle = \tfrac{1}{2} \sum_{i,j=1}^N x_i a_{ij} x_j - \sum_{i=1}^N b_i x_i,$$

with $A \in \mathbb{R}^{N \times N}$ symmetric and $b \in \mathbb{R}^N$.

1.3 Gradient and Hessian matrix

Definition 1.2 – (Differentiability and gradient)

Let $\mathcal{H}$ be a real Hilbert space, $U \subset \mathcal{H}$ an open set and $f : U \to \mathbb{R}$. The function f is *differentiable* (in the Fréchet sense) at $x \in U$ if there exists a continuous linear map $\mathrm{D}f(x) : \mathcal{H} \to \mathbb{R}$ such that

$$f(x+h) = f(x) + \mathrm{D}f(x)[h] + o(\|h\|), \qquad h \to 0. \tag{1}$$

By the Riesz representation theorem, there exists a unique $\nabla f(x) \in \mathcal{H}$ such that

$$\mathrm{D}f(x)[h] = \langle \nabla f(x) \,|\, h \rangle, \qquad \forall h \in \mathcal{H}, \tag{2}$$

called the *gradient* of f at x. When $\mathcal{H} = \mathbb{R}^N$ is endowed with the usual inner product, the gradient is the vector of partial derivatives: $\nabla f(x) = \left(\dfrac{\partial f(x)}{\partial x_i} \right)_{1 \le i \le N}$.

Remark (Gâteaux vs Fréchet) If only the directional limit $\lim_{t \to 0} \frac{f(x+th)-f(x)}{t} = \langle \nabla f(x) \,|\, h \rangle$ exists for every h, we speak of differentiability in the Gâteaux sense; this is the notion used for directional derivatives. Fréchet implies Gâteaux, the converse is false, but the two coincide as soon as the gradient is continuous: for the C^1 functions of this course, the distinction has no consequence.

Definition 1.3 – (Hessian matrix)

The second derivative of a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, or Hessian matrix of f at x, denoted $H_f(x) = \nabla^2 f(x) \in \mathbb{R}^{N \times N}$, is

$$H_f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \, \partial x_j} \right)_{1 \le i \le N,\, 1 \le j \le N}.$$

If f is of class C^2, the Hessian is symmetric (Schwarz's theorem).

Example 1.2 (Two computations to know)**.**

(a) Let $A \in \mathbb{R}^{N \times N}$ be symmetric, $z \in \mathbb{R}^N$ and $f(x) = \frac{1}{2} \langle x \,|\, Ax \rangle - \langle z \,|\, x \rangle$. Then

$$\nabla f(x) = Ax - z, \qquad H_f(x) = A.$$

(If A is not symmetric, $\nabla f(x) = \frac{1}{2}(A + A^\top)x - z$.)

(b) Let $A \in \mathbb{R}^{K \times N}$, $z \in \mathbb{R}^K$ and $f(x) = \frac{1}{2} \|Ax - z\|^2$. Then

$$\nabla f(x) = A^\top (Ax - z), \qquad H_f(x) = A^\top A.$$

1.4 Norms in finite dimension

Vectors. Let $x = (x_i)_{1 \le i \le N} \in \mathbb{R}^N$.

- ℓ_1 norm: $\|x\|_1 = \sum_i |x_i|$.
- ℓ_2 norm: $\|x\|_2 = \sqrt{\sum_i x_i^2}$.

Symmetric matrices. Let A be a real symmetric matrix of size $N \times N$.

- *Spectral decomposition*: A factorizes as

$$A = Q\Lambda Q^\top,$$

where $Q \in \mathbb{R}^{N\times N}$ is orthogonal ($Q^\top Q = \mathrm{Id}$) and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$, the real numbers λ_i being the eigenvalues of A. The columns of Q form an orthonormal basis of eigenvectors of A.

- *Spectral norm*: $\|A\|_2 = \max_i |\lambda_i|$.
- *Frobenius norm*: $\|A\|_F = \sqrt{\sum_i \lambda_i^2}$ $(= \sqrt{\sum_{i,j} a_{ij}^2})$.
- *Positive semidefinite / positive definite matrix*:
  - A is *positive semidefinite*, denoted $A \succeq 0$, if $x^\top Ax \geq 0$ for every $x \in \mathbb{R}^N$;
  - A is *positive definite*, denoted $A \succ 0$, if $x^\top Ax > 0$ for every $x \in \mathbb{R}^N \setminus \{0\}$.
  - We write $A \succeq B$ if $A - B \succeq 0$.
- By the spectral decomposition, setting $y = Q^\top x$, we have $x^\top Ax = \sum_{i=1}^N \lambda_i y_i^2$, hence

$$A \succeq 0 \iff \lambda_{\min}(A) \geq 0, \qquad A \succ 0 \iff \lambda_{\min}(A) > 0.$$

In the latter case, $x^\top Ax \geq \lambda_{\min}(A)\,\|x\|^2$ for every x, and A is invertible.

General matrices. Let $A \in \mathbb{R}^{K\times N}$, not necessarily square or symmetric, of rank $r \leq \min(K, N)$.

- *Singular value decomposition (SVD)*: A factorizes as

$$A = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{K\times K}$ and $V \in \mathbb{R}^{N\times N}$ are orthogonal ($U^\top U = \mathrm{Id}$, $V^\top V = \mathrm{Id}$) and $\Sigma \in \mathbb{R}^{K\times N}$ is "diagonal", with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ (followed by zeros). The σ_i are the *singular values* of A. The columns of V (resp. U) are orthonormal eigenvectors of $A^\top A$ (resp. $AA^\top$), and

$$\sigma_i = \sqrt{\lambda_i(A^\top A)} = \sqrt{\lambda_i(AA^\top)}.$$

- *Spectral norm* (operator norm induced by $\|\cdot\|_2$):

$$\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}.$$

We have $\left\|A^\top\right\|_2 = \|A\|_2$ and $\left\|A^\top A\right\|_2 = \|A\|_2^2$.

- *Frobenius norm*:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)} = \sqrt{\sum_{i=1}^r \sigma_i^2},$$

and $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\,\|A\|_2$.

- *Condition number* (if A has full rank $r = N$): $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$, with $\kappa(A^\top A) = \kappa(A)^2$.

1.5 Linear operators, norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

Definition 1.4 – (Bounded linear operator)

An operator $L : \mathcal{H} \to \mathcal{G}$ is *linear* if

$$(\forall a, b \in \mathbb{R})(\forall x, y \in \mathcal{H}) \qquad L(ax + by) = aL(x) + bL(y).$$

A linear operator L is *bounded* (or continuous) if

$$\|L\| = \sup_{\|x\|_{\mathcal{H}} \leq 1} \|L(x)\|_{\mathcal{G}} < +\infty.$$

We denote by $\mathcal{B}(\mathcal{H}, \mathcal{G})$ the Banach space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$.

Remark In finite dimension, every linear operator is bounded.

Definition 1.5 – (Adjoint)

Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Its adjoint L^* is the operator in $\mathcal{B}(\mathcal{G}, \mathcal{H})$ defined by

$$\big(\forall (x, y) \in \mathcal{H} \times \mathcal{G}\big) \qquad \langle y \,|\, L(x) \rangle_{\mathcal{G}} = \langle L^*(y) \,|\, x \rangle_{\mathcal{H}} \,.$$

Example 1.3 (Duplication operator)**.** If $L : \mathcal{H} \to \mathcal{H}^n$, $x \mapsto (x, \ldots, x)$, then

$$L^* : \mathcal{H}^n \to \mathcal{H} : y = (y_1, \ldots, y_n) \mapsto \sum_{i=1}^{n} y_i.$$

Proof. $\langle L(x) \,|\, y \rangle = \langle (x, \ldots, x) \,|\, (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} \langle x \,|\, y_i \rangle = \langle x \,|\, \sum_{i=1}^{n} y_i \rangle$.

A few remarks about L^*:

- It appears in the computation of gradients (see Example 1.2(b)) and of proximal operators, as well as in dual formulations.
- In finite dimension: if $L \in \mathcal{B}(\mathbb{R}^N, \mathbb{R}^K)$ is represented by a matrix, then $L^*(y) = L^\top y$.
- **Numerical check.** The implementation of the adjoint is checked by testing its definition on random vectors:

$$\big(\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^K\big) \qquad \langle L(x) \,|\, y \rangle = \langle x \,|\, L^*(y) \rangle \,.$$

A few remarks about $\|L\|$.

- This quantity is needed to set the step size of gradient-type algorithms (see Chapter 3).
- We have $\|L^*\| = \|L\|$.
- In practice, it is estimated by the power iteration method applied to L^*L, since $\|L\|^2 = \|L^*L\|$.

Power iteration method. To compute $\beta = \|L\|^2$ without forming or diagonalizing the matrix (often too large in imaging), we iterate from a random $x^{[0]}$

$$x^{[k+1]} = \frac{L^*L\, x^{[k]}}{\big\|L^*L\, x^{[k]}\big\|}, \qquad \rho_k = \big\|L^*L\, x^{[k]}\big\| \xrightarrow[k \to \infty]{} \lambda_{\max}(L^*L).$$

Only the functions $x \mapsto Lx$ and $y \mapsto L^*y$ are needed, which makes the method very easy to implement. One must first check that the function L^* is indeed the adjoint of L, with the test $\langle Lx \,|\, y \rangle = \langle x \,|\, L^*y \rangle$ on random vectors.

```python
import numpy as np

def power_iteration(L, Lt, N, n_iter=100, tol=1e-6, seed=0):
    """Estimate ||L||^2 = lambda_max(L^* L).
    L  : function x -> L(x)     (direct operator)
    Lt : function y -> L^*(y)   (adjoint operator)
    N  : dimension (or shape) of the input space
    """
    rng = np.random.default_rng(seed)
    x = rng.standard_normal(N)
    x /= np.linalg.norm(x)
    rho = 0.0
    for k in range(n_iter):
        y = Lt(L(x))                    # y = L^* L x
        rho_new = np.linalg.norm(y)     # estimate of lambda_max(L^* L)
        x = y / rho_new                 # normalization
        if abs(rho_new - rho) <= tol * rho_new:
            break
        rho = rho_new
    return rho_new                      # ||L||^2  (||L|| = sqrt(rho_new))

# Example 1: arbitrary matrix
A = np.random.default_rng(1).standard_normal((50, 30))
beta = power_iteration(lambda x: A @ x, lambda y: A.T @ y, 30)
print(beta, np.linalg.norm(A, 2)**2)          # identical values

# Example 2: periodic 2D discrete gradient (matrix-free)
def D(x):
    return np.stack([np.roll(x, -1, 0) - x, np.roll(x, -1, 1) - x])
def Dt(y):
    return (np.roll(y[0], 1, 0) - y[0]) + (np.roll(y[1], 1, 1) - y[1])
print(power_iteration(D, Dt, (64, 64), n_iter=500))   # close to 8
```

1.6 Approximation by Taylor expansion

Most optimization methods use polynomial approximations of the nonlinear objective function. The theoretical foundation is Taylor's theorem.

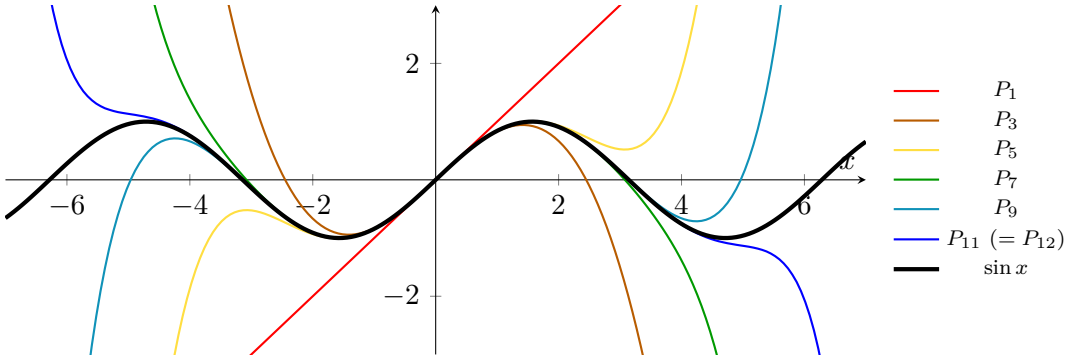


Figure 1: Taylor polynomials of sin at $a = 0$ of degrees $1, 3, 5, 7, 9, 11$ (the polynomial of degree 12 coincides with the one of degree 11, since the even-order derivatives of sin vanish at 0).

Theorem 1.6 – (Taylor, scalar case – Peano remainder)

Let $\ell \geq 1$ be an integer and $f : \mathbb{R} \to \mathbb{R}$ a function that is ℓ times differentiable at the point $a \in \mathbb{R}$. There exists a function $h_\ell : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(\ell)}(a)}{\ell!}(x-a)^\ell + h_\ell(x)(x-a)^\ell,$$
$$\text{with} \quad \lim_{x \to a} h_\ell(x) = 0.$$

This is the Peano form of the remainder.

The polynomial that appears is the *Taylor polynomial of order* ℓ of f at the point a:

$$P_\ell(x) = \sum_{i=0}^{\ell} \frac{f^{(i)}(a)}{i!}(x-a)^i.$$

It is the unique polynomial of "best asymptotic approximation": if there exist a function h_ℓ and a polynomial p of degree $\leq \ell$ such that $f(x) = p(x) + h_\ell(x)(x-a)^\ell$ with $\lim_{x\to a} h_\ell(x) = 0$, then $p = P_\ell$.

Generalization to dimension N. Let $f : \mathbb{R}^N \to \mathbb{R}$ and $h \in \mathbb{R}^N \setminus \{0\}$.

1. **Taylor polynomial of order 1** (best affine approximation of f in a neighborhood of x), for f differentiable:
$$T_1(x,h) = f(x) + \nabla f(x)^\top h. \tag{3}$$
2. **Taylor polynomial of order 2** (best quadratic approximation), for f of class C^2:
$$T_2(x,h) = f(x) + \nabla f(x)^\top h + \tfrac{1}{2} h^\top H_f(x)\, h. \tag{4}$$

In dimension N, one usually stops at order two: higher-order derivatives are difficult or expensive to evaluate and are not used in practice.

In proofs, we will use the *Lagrange form* (mean value form) of these expansions.

Proposition 1.7 – (Taylor formulas with Lagrange remainder)

Let $x, p \in \mathbb{R}^N$.

(i) If f is C^1 on a neighborhood of the segment $[x, x+p]$, there exists $t \in]0,1[$ such that
$$f(x+p) = f(x) + \nabla f(x+tp)^\top p. \tag{5}$$

(ii) If f is C^2 on a neighborhood of $[x, x+p]$, there exists $t \in]0,1[$ such that
$$f(x+p) = f(x) + \nabla f(x)^\top p + \tfrac{1}{2}\, p^\top H_f(x+tp)\, p. \tag{6}$$

These formulas are obtained by applying the scalar Taylor theorem to $\varphi(s) = f(x+sp)$ on $[0,1]$.

2 Classes of functions

Two main families of assumptions on f structure the course:

- **Regularity:** the class $C^{\ell,p}_\beta(\mathbb{R}^N)$: continuously differentiable functions with Lipschitz derivatives;

- **Convexity:** the class $\Gamma_0(\mathcal{H})$: proper, lower semicontinuous and convex functions.

2.1 Continuously differentiable functions

Definition 2.1 – (Classes of regular functions)

We follow here the notation of [1].

- $f \in C^\ell(\mathbb{R}^N)$ denotes the class of functions that are ℓ times continuously differentiable on $\mathbb{R}^N$.
- $f \in C^{\ell,p}_\beta(\mathbb{R}^N)$ denotes the class of functions that are ℓ times continuously differentiable on $\mathbb{R}^N$ and whose p-th derivative is Lipschitz continuous on $\mathbb{R}^N$ with constant β.
- In particular, $C^{1,1}_\beta(\mathbb{R}^N)$ is the class of functions with a β-Lipschitz gradient:

$$(\forall x, y \in \mathbb{R}^N) \qquad \|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\| .$$

- $f \in C^{2,1}_\beta(S)$, $S \subset \mathbb{R}^N$, if and only if $(\forall x \in S)\ \|\nabla^2 f(x)\| \leq \beta$.

Definition 2.2 – (β-smooth function)

A function $f \in C^{1,1}_\beta$ is said to be *Lipschitz smooth* with constant β, or β-*smooth*: its gradient is Lipschitz continuous with constant β,

$$(\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N) \qquad \|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| .$$

Remark $f = \frac{1}{2}\|A \cdot - z\|_2^2$ has a β-Lipschitz gradient with $\beta = \sigma_{\max}(A^\top A) = \|A\|^2$, since $\nabla^2 f = A^\top A$.

2.2 Domain, epigraph, lower semicontinuity

Definition 2.3 – (Domain, proper function, epigraph)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- The *domain* of f is $\operatorname{dom} f = \{x \in \mathcal{H} \mid f(x) < +\infty\}$.
- f is *proper* if $\operatorname{dom} f \neq \varnothing$.
- The *epigraph* of f is $\operatorname{epi} f = \{(x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \mid f(x) \leq \zeta\}$.

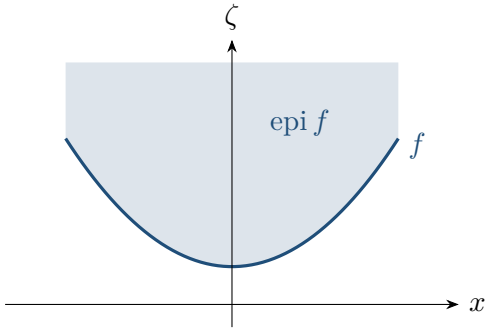


Figure 2: The epigraph of a function is the set of points lying above its graph.

Definition 2.4 – (Lower semicontinuity)

Let $f : \mathcal{H} \to]-\infty, +\infty]$. The function f is *lower semicontinuous* (l.s.c.) on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed.
Equivalently: for every $x \in \mathcal{H}$ and every sequence $x_n \to x$, $f(x) \leq \liminf_{n\to\infty} f(x_n)$.

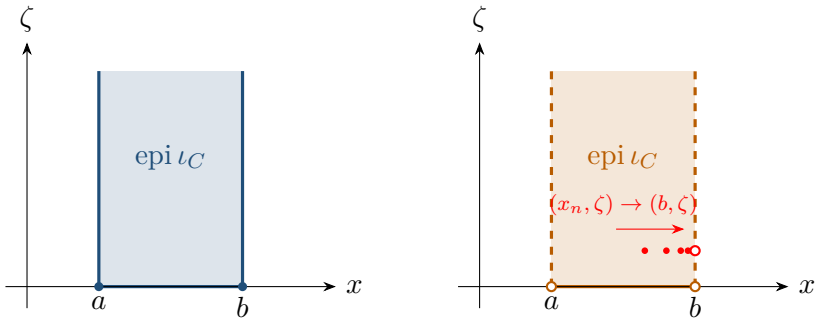


Figure 3: Epigraph of the indicator function ι_C: $\operatorname{epi} \iota_C = C \times [0, +\infty[$. Left: $C = [a, b]$ is closed, and so is the epigraph. Right: $C =]a, b[$ is open: the vertical edges (dashed) do not belong to the epigraph. A sequence of points (x_n, ζ) of the epigraph converges to $(b, \zeta) \notin \operatorname{epi} \iota_C$, so the epigraph is not closed. Indeed, $\iota_C(x_n) = 0$ whereas $\iota_C(b) = +\infty > \liminf \iota_C(x_n)$.

Examples (constraints). Through indicator functions (Definition 2.8), lower semicontinuity

- **does not allow** strict constraints, for example $Ax < b$ or $x > 0$ (open sets);
- **allows** equality or non-strict inequality constraints, for example $Ax = b$, $Ax \leq b$ or $x \geq 0$ (closed sets).

Properties.

- Every continuous function on $\mathcal{H}$ is l.s.c.
- Every finite sum of l.s.c. functions is l.s.c.

2.3 Convexity

Definition 2.5 – (Convex set)

$C \subset \mathcal{H}$ is *convex* if

$$(\forall (x, y) \in C^2)(\forall \alpha \in]0, 1[) \qquad \alpha x + (1 - \alpha) y \in C.$$

Definition 2.6 – (Convex function)

$f : \mathcal{H} \to]-\infty, +\infty]$ is *convex* if

$$(\forall (x, y) \in \mathcal{H}^2)(\forall \alpha \in]0, 1[) \qquad f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y).$$

Equivalently, f is convex if and only if its epigraph is convex.

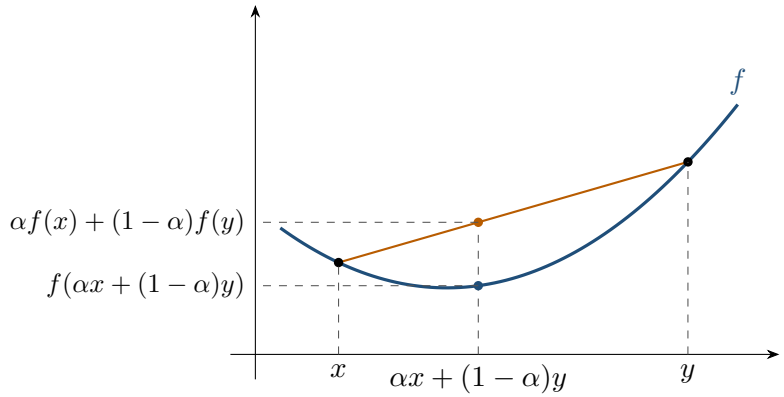


Figure 4: Convexity: the chord joining $(x, f(x))$ and $(y, f(y))$ lies above the graph.

Properties.

- The composition of a nondecreasing convex function with a convex function is convex.
- If $f : \mathcal{H} \to]-\infty, +\infty]$ is convex, then $\operatorname{dom} f$ is convex.
- $f : \mathcal{H} \to [-\infty, +\infty[$ is *concave* if $-f$ is convex.
- Every finite sum of convex functions is convex.

Definition 2.7 – (Class $\Gamma_0(\mathcal{H})$)

We denote by $\Gamma_0(\mathcal{H})$ the set of functions $f : \mathcal{H} \to]-\infty, +\infty]$ that are proper, lower semicontinuous and convex.

Definition 2.8 – (Indicator function)

Let $C \subset \mathcal{H}$. The indicator function of C is

$$(\forall x \in \mathcal{H}) \qquad \iota_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

We have: $\iota_C \in \Gamma_0(\mathcal{H}) \iff C$ is a nonempty closed convex set.

Thanks to the indicator function, a constrained problem $\min_{x \in C} f(x)$ can be rewritten as the unconstrained problem $\min_{x \in \mathcal{H}} f(x) + \iota_C(x)$.

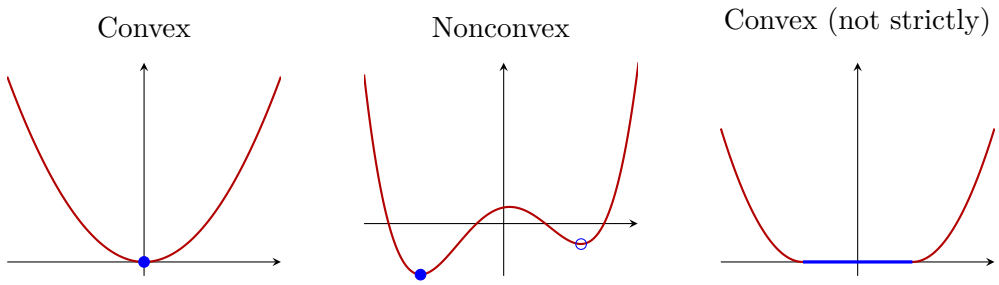


Figure 5: Convex function (unique minimizer), nonconvex function (one global minimizer • and one local minimizer ∘), convex but not strictly convex function (the set of minimizers is not a single point).

2.4 Strict and strong convexity

Definition 2.9 – (Strict convexity)

Let $f : \mathcal{H} \to]-\infty, +\infty]$. f is *strictly convex* if

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[) \qquad x \neq y \implies f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

Definition 2.10 – (Strongly convex function)

Let $\mu > 0$. A function $f : \mathcal{H} \to]-\infty, +\infty]$ is *μ-strongly convex* if

$$f - \frac{\mu}{2} \|\cdot\|^2 \text{ is convex.}$$

Equivalently:

(i) $(\forall x, y \in \mathcal{H})(\forall t \in [0,1])$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\mu}{2} t(1-t) \|x - y\|^2 ;$$

(ii) if f is differentiable: $(\forall x, y \in \mathcal{H}) \quad f(y) \geq f(x) + \langle \nabla f(x) \,|\, y - x \rangle + \frac{\mu}{2} \|y - x\|^2$;

(iii) if f is of class C^2: $(\forall x) \quad \nabla^2 f(x) \succeq \mu \operatorname{Id}$.

Remark

- A strongly convex function is strictly convex, but the converse is false (for instance $x \mapsto x^4$ or $x \mapsto e^x$ on $\mathbb{R}$).

2.5 Characterizations of convexity

Theorem 2.11 – (Convexity and derivatives)

Let C be a convex (open) subset of $\mathbb{R}^N$ and $f : C \to \mathbb{R}$ twice differentiable on C. The following statements are equivalent:

(i) f is convex;

(ii) $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in C$;

(iii) $\nabla^2 f(x) \succeq 0$ for all $x \in C$.

Moreover:

(iv) f is strictly convex if and only if $f(y) > f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in C$, $x \neq y$;

(v) $\nabla^2 f(x) \succ 0$ for all $x \in C$ implies that f is strictly convex (the converse is false: $x \mapsto x^4$).

Example 2.1 (Back to quadratic functions)**.**

For $f(x) = \frac{1}{2} x^\top A x - b^\top x$ with A symmetric (Example 1.1):

- gradient: $\nabla f(x) = Ax - b$; Hessian: $H_f(x) = A$;
- f is called *positive definite* if A is positive definite;
- a positive definite quadratic function is strictly convex (and even $\lambda_{\min}(A)$-strongly convex).

2.6 Weak convexity

Between convex functions and arbitrary functions, the class of *weakly convex* functions gathers the functions that are "convex up to a quadratic term".

Definition 2.12 – (Weakly convex function)

Let $\rho \geq 0$. A function $f : \mathcal{H} \to]-\infty, +\infty]$ is *ρ-weakly convex* if

$$f + \frac{\rho}{2}\|\cdot\|^2 \text{ is convex.}$$

Equivalently:

(i) $(\forall x, y \in \mathcal{H})(\forall t \in [0,1])$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) + \frac{\rho}{2}t(1-t)\|x-y\|^2;$$

(ii) if f is differentiable: $(\forall x, y \in \mathcal{H}) \quad f(y) \geq f(x) + \langle \nabla f(x) \,|\, y - x\rangle - \frac{\rho}{2}\|y-x\|^2$;

(iii) if f is of class C^2: $(\forall x) \quad \nabla^2 f(x) \succeq -\rho\,\mathrm{Id}$.

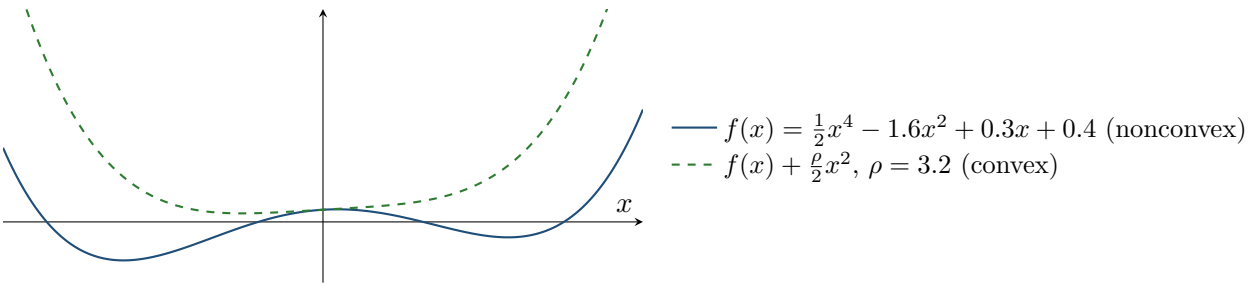


Figure 6: A nonconvex but ρ-weakly convex function: adding the quadratic term $\frac{\rho}{2}x^2$ makes it convex.

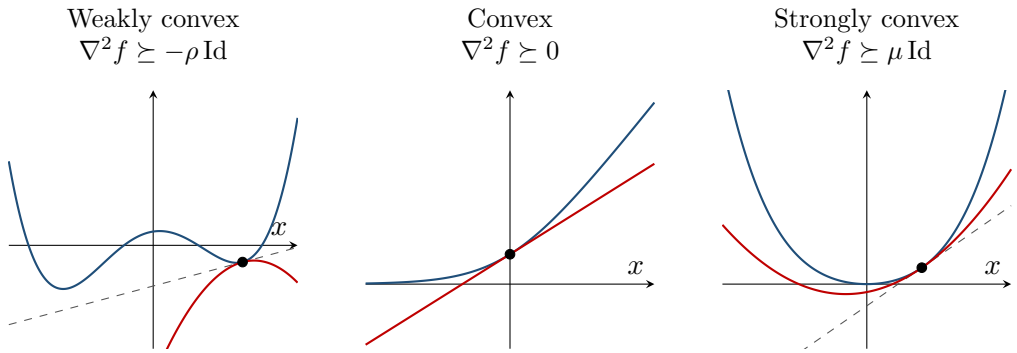


Figure 7: One and the same first-order inequality for the three notions: $f(y) \geq f(x) + \langle \nabla f(x) \,|\, y - x\rangle + \frac{c}{2}\|y-x\|^2$ at every point x (•), with $c = -\rho < 0$ (weak convexity), $c = 0$ (convexity) and $c = \mu > 0$ (strong convexity). In blue the function f, in red the quadratic minorant, dashed the tangent. Left: f goes below its tangent (it is not convex) but stays above a concave parabola ($\rho = 3.2$); middle: the tangent is enough; right: f stays above a convex parabola ($\mu = 1$).

Proposition 2.13 – (Convex criterion with a nonconvex penalty)

Let φ be a ρ-weakly convex function and $\lambda > 0$. The denoising criterion

$$F(x) = \tfrac{1}{2}\|x - z\|^2 + \lambda\,\varphi(x)$$

is convex as soon as $\lambda\rho \leq 1$, and $(1 - \lambda\rho)$-strongly convex as soon as $\lambda\rho < 1$.

Proof. We write $F(x) = \lambda(\varphi(x) + \frac{\rho}{2}\|x\|^2) + \frac{1-\lambda\rho}{2}\|x\|^2 - \langle z \,|\, x\rangle + \frac{1}{2}\|z\|^2$. The first term is convex by definition, the second one is $(1 - \lambda\rho)$-strongly convex if $\lambda\rho < 1$ (convex if $\lambda\rho = 1$), and the remaining ones are affine. ■

Remark This is the principle of "nonconvex penalties leading to a convex criterion" (Parekh, Selesnick, 2015): one benefits from the lower bias of a nonconvex penalty (MCP, logarithmic penalty) while keeping a convex problem.

3 Minimizers and optimality conditions

We seek $\widehat{x} \in \operatorname{Argmin}_{x\in C} f(x)$. Three questions arise: local versus global minimizers; existence (through coercivity); the role of convexity.

3.1 Local and global minimizers, stationary points

Definition 3.1 – (Minimizers)

Let C be a nonempty subset of a Hilbert space $\mathcal{H}$, $f : C \to]-\infty, +\infty]$ a proper function and $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a *local minimizer* of f if there exists an open neighborhood O of $\widehat{x}$ such that $(\forall x \in O \cap C)\ f(\widehat{x}) \leq f(x)$.
- $\widehat{x}$ is a *(global) minimizer* of f if $(\forall x \in C)\ f(\widehat{x}) \leq f(x)$.
- $\widehat{x}$ is a *strict local minimizer* if there exists an open neighborhood O of $\widehat{x}$ such that $(\forall x \in (O \cap C) \setminus \{\widehat{x}\})\ f(\widehat{x}) < f(x)$.
- $\widehat{x}$ is a *strict global minimizer* if $(\forall x \in C \setminus \{\widehat{x}\})\ f(\widehat{x}) < f(x)$.
- If f is differentiable at $\widehat{x} \in \operatorname{dom} f$, $\widehat{x}$ is a *stationary* (or critical) *point* of f if $\nabla f(\widehat{x}) = 0$.

3.2 How to recognize a minimum?

- **Necessary conditions** for a point to be a minimum: first-order conditions (they only use the gradient).
- **Sufficient conditions**: second-order conditions (they use the gradient and the Hessian matrix).

Theorem 3.2 – (First-order necessary condition)

Let $f \in C^1(\Omega)$ on a neighborhood Ω of $\widehat{x}$. If $\widehat{x}$ is a minimizer of f (in Ω), then $\nabla f(\widehat{x}) = 0$, that is, $\widehat{x}$ is a stationary point of f.

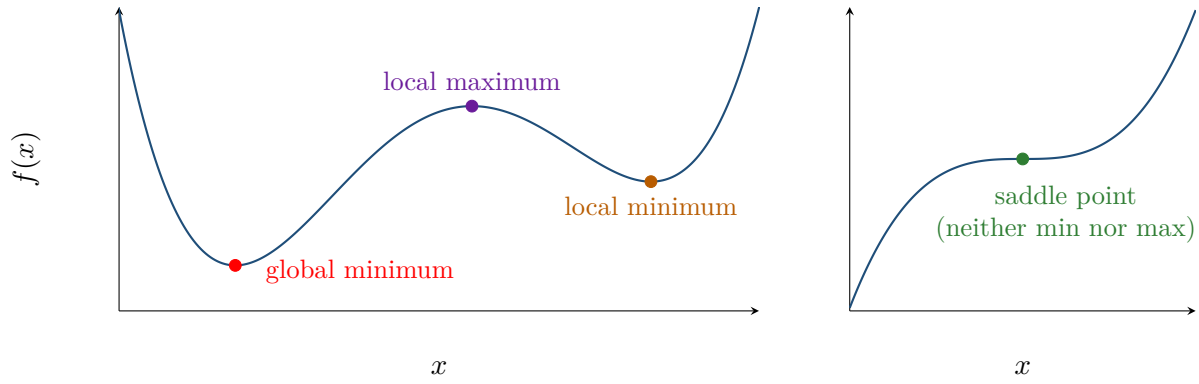


Figure 8: Stationary points: a point where the gradient vanishes can be a global minimum, a local minimum, a local maximum (left) or a saddle / inflection point, such as $x = 0$ for $x \mapsto x^3$ (right).

Proof. By contradiction, assume $\nabla f(\widehat{x}) \neq 0$. Let $p = -\nabla f(\widehat{x})$ be the anti-gradient of f at $\widehat{x}$; clearly $p \neq 0$. The function $g(x) = \nabla f(x)^\top p$ satisfies

$$g(\widehat{x}) = \nabla f(\widehat{x})^\top p = -\nabla f(\widehat{x})^\top \nabla f(\widehat{x}) = -\left\|\nabla f(\widehat{x})\right\|^2 < 0.$$

Since $f \in C^1(\Omega)$, g is continuous on Ω and therefore remains negative in a neighborhood of $\widehat{x}$: there exists $T > 0$ such that

$$(\forall t \in [0, T]) \qquad 0 > g(\widehat{x} + tp) = \nabla f(\widehat{x} + tp)^\top p. \tag{7}$$

By the mean value theorem, for every $\tau \in]0, T[$, there exists $t \in]0, 1[$ such that, setting $t' = t\tau \in]0, T[$,

$$f(\widehat{x} + \tau p) = f(\widehat{x}) + \nabla f(\widehat{x} + t'p)^\top \tau p = f(\widehat{x}) + \tau \underbrace{\nabla f(\widehat{x} + t'p)^\top p}_{<0 \text{ by } (7)} < f(\widehat{x}).$$

Hence $\widehat{x}$ cannot be a minimum of f: contradiction. ■

Caution This is a *necessary* condition: all minimizers are stationary points, but not all stationary points are minimizers; they can be maximizers or saddle points (Figure 8).

Theorem 3.3 – (Second-order necessary condition)

Let $f \in C^2(\Omega)$ on a neighborhood Ω of $\widehat{x}$. If $\widehat{x}$ is a minimizer of f (in Ω), then $H_f(\widehat{x})$ is positive semidefinite.

Proof. By contradiction, assume that $H_f(\widehat{x})$ is not positive semidefinite: there exists $p \in \mathbb{R}^N$, $p \neq 0$, such that $p^\top H_f(\widehat{x}) p < 0$. Set $g(x) = p^\top H_f(x) p$; we have $g(\widehat{x}) < 0$. Since $f \in C^2(\Omega)$, g is continuous on Ω and there exists $T > 0$ such that

$$(\forall t \in [0, T]) \qquad 0 > g(\widehat{x} + tp) = p^\top H_f(\widehat{x} + tp)\, p. \tag{8}$$

By the mean value formula (Lagrange remainder), for every $\tau \in]0, T[$, there exists $t \in]0, 1[$ such that, with $t' = t\tau \in]0, T[$,

$$f(\widehat{x}+\tau p) = f(\widehat{x}) + \underbrace{\nabla f(\widehat{x})^\top}_{=0 \text{ (Theorem 3.2)}} \tau p + \tfrac{1}{2}(\tau p)^\top H_f(\widehat{x}+t'p)(\tau p) = f(\widehat{x}) + \tfrac{1}{2}\tau^2 \underbrace{p^\top H_f(\widehat{x} + t'p) p}_{<0 \text{ by } (8)} < f(\widehat{x}).$$

Hence $\widehat{x}$ cannot be a minimum of f: contradiction. ■

Theorem 3.4 – (Second-order sufficient condition)

Let $f \in C^2(\Omega)$ on a neighborhood Ω of $\widehat{x}$. Then

$$\left.\begin{array}{l}\nabla f(\widehat{x}) = 0 \\ H_f(\widehat{x}) \text{ positive definite}\end{array}\right\} \Longrightarrow \widehat{x} \text{ is a (strict) local minimizer of } f.$$

Proof. Since $H_f(\widehat{x})$ is positive definite and H_f is continuous, there exists a ball $B = B_T(\widehat{x})$ centered at $\widehat{x}$ such that $H_f(x)$ remains positive definite for every $x \in B$. Then, for every $p \in \mathbb{R}^N$ with $\|p\| = 1$ and every $\tau \in]0, T[$, there exists $t \in]0, 1[$ such that, with $t' = t\tau$,

$$f(\widehat{x}+\tau p) = f(\widehat{x}) + \underbrace{\nabla f(\widehat{x})^\top}_{=0}\tau p + \tfrac{1}{2}(\tau p)^\top H_f(\widehat{x}+t'p)(\tau p) = f(\widehat{x}) + \underbrace{\tfrac{1}{2}\tau^2\, p^\top H_f(\underbrace{\widehat{x}+t'p}_{\in B})p}_{>0} > f(\widehat{x}).$$

Thus $f(\widehat{x})$ is the smallest value taken by f on B: $\widehat{x}$ is a (strict local) minimizer of f. ■

Remark There is no sufficient condition that uses first derivatives only (in the absence of convexity). Moreover, checking that $H_f(\widehat{x})$ is positive definite is generally expensive (computation of the eigenvalues); this condition is therefore rarely used in practice.

3.3 The convex case

Theorem 3.5 – (Minimizers of a convex function)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper convex function such that $\mu = \inf f > -\infty$.

- The set $\{x \in \mathcal{H} \mid f(x) = \mu\}$ of minimizers is convex.
- Every local minimizer of f is a global minimizer.
- If f is strictly convex, f has at most one minimizer.

Definition 3.6 – (Coercivity)

$f : \mathcal{H} \to]-\infty, +\infty]$ is *coercive* if $\lim\limits_{\|x\|\to+\infty} f(x) = +\infty$.

Theorem 3.7 – (Existence and uniqueness of a minimizer)

Let C be a closed convex subset of $\mathcal{H}$ and $f \in \Gamma_0(\mathcal{H})$ such that $\operatorname{dom} f \cap C \neq \varnothing$. If f is coercive or if C is bounded, then there exists $\widehat{x} \in C$ such that

$$f(\widehat{x}) = \inf_{x\in C} f(x).$$

If, in addition, f is strictly convex, this minimizer $\widehat{x}$ is unique.

Theorem 3.8 – (Optimality condition (Fermat's rule))

Let $f \in \Gamma_0(\mathbb{R}^N) \cap C^1(\mathbb{R}^N)$. Then

$$\widehat{x} \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} f(x) \quad \Longleftrightarrow \quad \nabla f(\widehat{x}) = 0.$$

In the convex case, the first-order condition is therefore both *necessary and sufficient*: this follows from Theorem 2.11(ii) with $x = \widehat{x}$.

Limitations of Fermat's rule.

- It leads to a system of N equations in N unknowns.
- An explicit expression (closed form) exists only in a few cases.
- Without a closed form, an *iterative* procedure is needed.

3.4 Examples

Theorem 3.9 – (Minimizer of a positive definite quadratic)

A positive definite quadratic function $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ has a unique minimizer $\widehat{x}$, which is the unique solution of the linear system

$$Ax = b.$$

Proof. f is strictly convex, hence has at most one minimizer. Its Hessian A is positive definite and, by Fermat's rule,

$$\widehat{x} \text{ minimizes } f \iff \nabla f(\widehat{x}) = 0 \iff A\widehat{x} = b.$$

Since A is positive definite, it is invertible: the system $Ax = b$ has a unique solution. ■

Example 3.1 (Least squares)**.** Let $A \in \mathbb{R}^{M\times N}$ have full rank ($\operatorname{rank} A = N$) and $y \in \mathbb{R}^M$. We seek $\widehat{x} = \arg\min_{x\in\mathbb{R}^N} \|Ax - y\|_2^2$. The optimality condition reads

$$\nabla f(\widehat{x}) = 0 \iff A^\top(A\widehat{x} - y) = 0 \iff \widehat{x} = (A^\top A)^{-1}A^\top y.$$

A closed form is available, but inverting $A^\top A$ can be difficult (large dimension, ill-conditioning): an iterative method is often preferred.

Example 3.2 (Logistic criterion)**.** Let $y \in \mathbb{R}$ and $f(x) = \log(1 + \exp(-yx))$, $x \in \mathbb{R}$. The optimality condition reads

$$\nabla f(\widehat{x}) = 0 \iff \frac{-y\exp(-y\widehat{x})}{1 + \exp(-y\widehat{x})} = 0.$$

There is no explicit expression: an iterative procedure is needed.

> **Remark** For a single data point $y \neq 0$, this criterion is strictly monotone and actually has *no* minimizer ($\inf f = 0$ is not attained: the function is not coercive). In practice, one considers regularized logistic regression $\min_{x\in\mathbb{R}^N} \sum_{i=1}^M \log\big(1 + \exp(-y_i\, a_i^\top x)\big) + \frac{\lambda}{2}\|x\|^2$, which is strongly convex and has a unique minimizer, but no closed form.

4 Iterative descent methods

Definition 4.1 – (Iterative scheme)

Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^N$, and $x^{[0]} \in \mathbb{R}^N$. We consider

$$(\forall k \in \mathbb{N}) \qquad x^{[k+1]} = x^{[k]} + \gamma_k\, d^{[k]}, \qquad \gamma_k > 0,$$

where $d^{[k]} \in \mathbb{R}^N$ is a *direction* and γ_k a *step size*.

An iterative descent method consists in building a sequence $(x^{[k]})_{k\in\mathbb{N}}$ such that, at each iteration,

$$f(x^{[k+1]}) = f(x^{[k]} + \gamma_k d^{[k]}) < f(x^{[k]}).$$

Key questions:

- How to choose the direction $d^{[k]}$?
- How to prove the convergence of $(x^{[k]})_{k\in\mathbb{N}}$ to $\widehat{x} \in \operatorname{Argmin} f$?
- How to choose γ_k to guarantee convergence? To accelerate convergence?

4.1 Descent directions

Definition 4.2 – (Directional derivative, descent direction)

Let $f : \mathbb{R}^N \to]-\infty, +\infty]$ be proper and differentiable in a neighborhood of $x \in \mathbb{R}^N$. The directional derivative of f at x in the direction $d \in \mathbb{R}^N$ is

$$\langle \nabla f(x) \,|\, d \rangle = \lim_{\gamma \to 0} \frac{f(x+\gamma d) - f(x)}{\gamma}.$$

A direction $d \in \mathbb{R}^N$ is a *descent direction* for f at x if

$$\langle \nabla f(x) \,|\, d \rangle < 0,$$

that is, if the angle ϑ between d and $\nabla f(x)$ satisfies $\vartheta \in]\frac{\pi}{2}, \pi]$.

Proposition 4.3

If d is a descent direction for f at x, then there exists $\bar{\gamma} > 0$ such that

$$f(x+\gamma d) - f(x) < 0 \qquad \forall \gamma \in]0, \bar{\gamma}[.$$

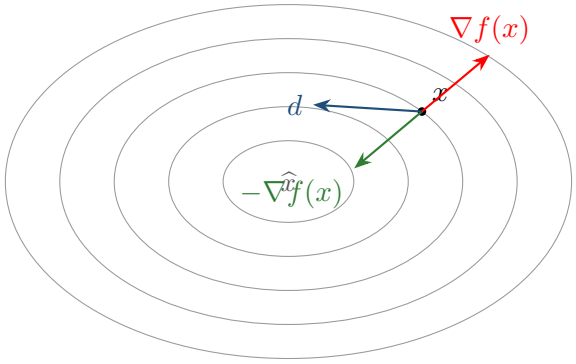


Figure 9: Level lines of a convex function, gradient and descent directions.

Example: steepest descent. If $\nabla f(x^{[k]}) \neq 0$, one can always find a descent direction: the anti-gradient

$$d^{[k]} = -\nabla f(x^{[k]}).$$

The steepest descent direction minimizes the directional derivative of f at $x^{[k]}$:

$$\nabla f(x^{[k]})^\top d = \left\| \nabla f(x^{[k]}) \right\| \, \|d\| \cos\vartheta,$$

where $\vartheta \in [0, \pi]$ is the angle between $\nabla f(x^{[k]})$ and d. Assuming without loss of generality that $\|d\| = 1$, the direction that maximizes the decrease minimizes $\cos\vartheta$, hence $\vartheta = \pi$ and

$$d^{[k]} = -\frac{\nabla f(x^{[k]})}{\|\nabla f(x^{[k]})\|}.$$

4.2 What do we expect from an optimization method?

- **Convergence to a stationary point**:

$$\lim_{k\to\infty} \left\|\nabla f(x^{[k]})\right\| = 0. \tag{9}$$

- **Global convergence**: (9) is guaranteed for every initialization $x^{[0]}$, regardless of how close it is to $\widehat{x}$. This is crucial for nonconvex problems, and it depends on the choice of the step size (*globalization strategies*).
- **Convergence of the sequence** (stronger): $\lim_{k\to\infty} x^{[k]} = \widehat{x}$. It is generally hard to establish, and often obtained only for initializations sufficiently close to the solution (*local convergence*). It is crucial to ensure convergence to local minima rather than to saddle points.

> **Caution** **Very important!** *Global* convergence does not mean convergence to a *global* minimum: it means convergence (to a stationary point) whatever the initialization.

5 Concrete examples: inverse problems, segmentation, learning

This section shows how the notions of the previous sections (gradient, Lipschitz constant β, strong convexity, Fermat's rule, convergence rate of gradient descent) apply to three families of problems encountered in imaging and learning. For each example, we systematically ask the same questions:

(a) what is the criterion f to be minimized?

(b) is f convex? strongly convex (and with which μ)?

(c) is f β-smooth (and with which β)?

(d) is there a closed form? if not, what does gradient descent guarantee?

5.1 Inverse problems in imaging

Notation. An image $x \in \mathbb{R}^{N_1\times N_2}$, $x = (x_{n_1,n_2})_{1\le n_1\le N_1, 1\le n_2\le N_2}$, is stacked column by column into a vector $x = (x_n)_{1\le n\le N} \in \mathbb{R}^N$ with $N = N_1 \times N_2$.

Forward model. We observe

$$z = \Phi x + w \in \mathbb{R}^M,$$

where $\Phi \in \mathbb{R}^{M\times N}$ models the acquisition process and w is noise.

Goal: build an estimate $\widehat{x} \in \mathbb{R}^N$ of x from z.

In general: Φ is known (or can be approximated) and sparse, but it is **ill-conditioned**.

Application	Operator Φ
Deblurring (convolution)	$\Phi x = \varphi * x$, convolution with the impulse response (PSF) φ of size $Q_1 \times Q_2$. Φ is block-circulant with circulant blocks: $\Phi = F^* \Lambda F$, with F the discrete Fourier transform ($F^* = F^{-1}$) and Λ diagonal.
MRI	$\Phi = MF$: 2D Fourier transform followed by a mask M that selects the measured Fourier coefficients (subsampling).
Tomography	$\Phi_{m,n}$: probability of detecting an event emitted at pixel n in the line of response m; z is the sinogram.

Table 1: Examples of forward models $z \simeq \Phi x$.

Remark (link with gradient descent, Chapter 3) For a convolution, $\Phi = F^* \Lambda F$ directly gives $\|\Phi\| = \max_n |\Lambda_{nn}|$ and $\Phi^\top \Phi = F^* |\Lambda|^2 F$: computing Φx, $\Phi^* z$ and the constant $\beta = \|\Phi\|^2$ only requires FFTs.

Ill-posed problem and inverse filtering

Definition 5.1 – (Hadamard conditions (1902))

The problem $z = \Phi x$ is *well-posed* if:

1. *existence*: $\operatorname{ran} \Phi = \mathbb{R}^M$ (every z is the image of some x);
2. *uniqueness*: $\ker \Phi = \{0\}$ (if $\widehat{x}(z)$ and $\widehat{x}'(z)$ are two solutions, $\widehat{x}(z) - \widehat{x}'(z) \in \ker \Phi$);
3. *stability*: $\|z - z'\| \to 0 \Rightarrow \|\widehat{x}(z) - \widehat{x}(z')\| \to 0$ (a small perturbation of the observation only causes a small change in the solution).

- If $M = N$ and Φ is invertible, *inverse filtering* gives $\widehat{x} = \Phi^{-1} z = x + \Phi^{-1} w$.
- If $M \geq N$ and $\operatorname{rank} \Phi = N$, this is the least-squares solution (see Section 3.4): $\widehat{x} = (\Phi^* \Phi)^{-1} \Phi^* z = x + (\Phi^* \Phi)^{-1} \Phi^* w$.

In both cases the error satisfies $\|\widehat{x} - x\| \leq \|w\| / \sigma_{\min}(\Phi)$: if Φ is ill-conditioned, the noise is **strongly amplified**. In optimization terms, $f(x) = \frac{1}{2} \|\Phi x - z\|^2$ is μ-strongly convex with $\mu = \sigma_{\min}(\Phi)^2$ and β-smooth with $\beta = \sigma_{\max}(\Phi)^2$: the condition number $\kappa = \beta/\mu$ is huge, which makes the solution unstable (and, as we will see in Chapter 3, gradient descent slow).

Regularization and Bayesian interpretation

The variational approach consists in seeking

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \underbrace{\tfrac{1}{2} \|\Phi x - z\|_2^2}_{\text{data fidelity}} + \lambda \underbrace{R(x)}_{\text{penalty}}, \tag{10}$$

where $\lambda > 0$ is the regularization parameter ($\lambda = 0$ gives back inverse filtering).

Maximum a posteriori (MAP). If x and z are realizations of random vectors X and Z, Bayes' rule gives

$$\max_x \mu_{X|Z=z}(x) \iff \max_x \mu_{Z|X=x}(z)\, \mu_X(x) \iff \min_x \underbrace{-\log \mu_{Z|X=x}(z)}_{L(x)\text{: data fidelity}} \underbrace{-\log \mu_X(x)}_{R(x)\text{: prior}}.$$

- **Gaussian noise** $w \sim \mathcal{N}(0, \alpha \mathrm{Id})$: $L(x) = \sum_{n=1}^{M} \frac{1}{2\alpha}\big((\Phi x)_n - z_n\big)^2$. It is a convex quadratic, with a β-Lipschitz gradient, $\beta = \|\Phi\|^2/\alpha$.
- **Poisson noise** $z = \mathcal{D}_\alpha(\Phi x)$: $L(x) = \sum_{n=1}^{M} \Psi_n((\Phi x)_n)$ with $\Psi_n(\upsilon) = \alpha\upsilon - z_n \ln(\alpha\upsilon)$ if $z_n > 0$ and $\upsilon > 0$, $\Psi_n(\upsilon) = \alpha\upsilon$ if $z_n = 0$ and $\upsilon \geq 0$, $+\infty$ otherwise. This function belongs to Γ_0 (convex, proper, l.s.c.), but **does not have a β-Lipschitz gradient**: $\Psi_n''(\upsilon) = z_n/\upsilon^2$ blows up when $\upsilon \to 0^+$. Fixed-step gradient descent therefore does not apply directly; this is an example where proximal algorithms are needed.

Tikhonov regularization (smooth case)

With $R(x) = \frac{1}{2}\|Dx\|^2$, where D is a finite-difference operator, the criterion

$$f(x) = \tfrac{1}{2}\|\Phi x - z\|^2 + \tfrac{\lambda}{2}\|Dx\|^2$$

is a quadratic function (Example 1.1):

- $\nabla f(x) = \Phi^\top(\Phi x - z) + \lambda D^\top D x$ and $\nabla^2 f = \Phi^\top\Phi + \lambda D^\top D$;
- f is β-smooth with $\beta = \lambda_{\max}(\Phi^\top\Phi + \lambda D^\top D) \leq \|\Phi\|^2 + \lambda\|D\|^2$;
- f is strongly convex as soon as $\ker\Phi \cap \ker D = \{0\}$ (for a blur, $\ker D$ = constant images, and Φ preserves the mean);
- Fermat's rule gives the system $(\Phi^\top\Phi + \lambda D^\top D)\widehat{x} = \Phi^\top z$ (Theorem 3.9); if Φ and D are circular convolutions, it is solved with one FFT.

Total variation (TV). For piecewise constant images, we penalize the discrete gradient $D = [D_1^\top, D_2^\top]^\top \in \mathbb{R}^{2N\times N}$ (horizontal differences $\psi_1 = [1\ -1]$ and vertical differences $\psi_2 = \psi_1^\top$):

- anisotropic TV [Rudin et al., 1992]:

$$R(x) = \|Dx\|_1 = \sum_{n_1,n_2} |x_{n_1+1,n_2} - x_{n_1,n_2}| + |x_{n_1,n_2+1} - x_{n_1,n_2}|\,;$$

- isotropic TV (coupling of the two directions):

$$R(x) = \|Dx\|_{1,2} = \sum_{n_1,n_2} \sqrt{|x_{n_1+1,n_2} - x_{n_1,n_2}|^2 + |x_{n_1,n_2+1} - x_{n_1,n_2}|^2}.$$

In 1D, the choice of the penalty $\|Dx\|_\bullet$ determines the nature of the solution: $\|\cdot\|_2^2$ (linear, smooth), $\|\cdot\|_1$ (piecewise constant), $\|\cdot\|_0$ (piecewise constant, nonconvex); with $\psi = [1\ -2\ 1]$ and $\|\cdot\|_1$, one obtains piecewise affine solutions.

Blind deconvolution. When Φ is unknown, we jointly estimate

$$(\widehat{x}, \widehat{\Phi}) \in \operatorname*{Argmin}_{x,\Phi} \tfrac{1}{2}\|\Phi x - z\|^2 + \lambda_1 R_1(x) + \lambda_2 R_2(\Phi).$$

The term Φx is bilinear in (x, Φ): the problem is **nonconvex** (even though each subproblem, with Φ or x fixed, is convex).

5.2 Image segmentation

Mumford–Shah model (1989). We seek a piecewise smooth approximation x of the image z and the set K of its contours:

$$\min_{x,K} \frac{1}{2}\int_\Omega (x-z)^2\,dx\,dy + \beta_{\mathrm{MS}}\int_{\Omega\setminus K}|\nabla x|^2\,dx\,dy + \lambda\,\mathcal{H}^1(K\cap\Omega),$$

(data fidelity, smoothness away from the contours, length of the contours, $\mathcal{H}^1$ being the Hausdorff measure).

Discrete Mumford–Shah. On the grid $\Omega = \{1,\dots,N_1\}\times\{1,\dots,N_2\}$, with $D\in\mathbb{R}^{|E|\times|\Omega|}$ a finite-difference operator and $e\in\mathbb{R}^{|E|}$ a contour variable on the edges (close to 1 where there is a contour, to 0 otherwise):

$$\min_{x,e}\ F(x,e) = \frac{1}{2}\|x-z\|_2^2 + \beta_{\mathrm{MS}}\,\|(1-e)\odot Dx\|^2 + \lambda R(e). \tag{11}$$

5.3 Supervised learning

We are given a database $S=\{(u_\ell,c_\ell)\in\mathcal{H}\times\mathcal{G}\mid \ell\in\{1,\dots,L\}\}$, for instance $u_\ell\in\mathbb{R}^N$ an image and $c_\ell\in\{-1,+1\}$ its class, and we seek a prediction function $d_\Theta:\mathcal{H}\to\mathcal{G}$.

- Choice of $\mathcal{H}$: $\mathbb{R}^N$ (image of size $N=N_1\times N_2$), $\mathbb{R}^{N\times M}$ (multivariate images);
- choice of $\mathcal{G}$: $\{-1,+1\}$ (binary classification), $\{1,\dots,K\}$ (multiclass), $\mathbb{R}$ (regression), $\mathbb{R}^K$ (multivariate regression).

Learning relies on a minimization problem

$$\widehat{\Theta}\in\operatorname*{Argmin}_\Theta\ \frac{1}{L}\sum_{\ell=1}^{L} F(c_\ell,d_\Theta(u_\ell)) + \lambda R(\Theta).$$

For a linear predictor $d_\Theta(u)=\Theta^\top u$ (or $\Theta^\top\phi(u)$ with a representation ϕ, for instance a *scattering* network), we write $U=[u_1,\dots,u_L]^\top\in\mathbb{R}^{L\times N}$.

Ridge regression. $f(\Theta)=\frac{1}{L}\sum_\ell(c_\ell-\Theta^\top u_\ell)^2+\lambda\|\Theta\|_2^2=\frac{1}{L}\|U\Theta-c\|^2+\lambda\|\Theta\|^2$. It is a quadratic:

$$\nabla f(\Theta)=\tfrac{2}{L}U^\top(U\Theta-c)+2\lambda\Theta,\qquad \nabla^2 f=\tfrac{2}{L}U^\top U+2\lambda\mathrm{Id},$$

hence $\mu=2\lambda$, $\beta=\frac{2}{L}\|U\|^2+2\lambda$, and Fermat's rule gives $\widehat{\Theta}=(U^\top U+\lambda L\,\mathrm{Id})^{-1}U^\top c$. The term $\lambda\|\Theta\|^2$ makes the problem strongly convex even if $U^\top U$ is not invertible ($L<N$): this is exactly the role of regularization in inverse problems.

Regularized logistic classification.

$$f(\Theta)=\frac{1}{L}\sum_{\ell=1}^{L}\log\!\left(1+e^{-c_\ell\Theta^\top u_\ell}\right)+\lambda\|\Theta\|_2^2.$$

With $s(t)=\log(1+e^{-t})$, we have $s'(t)=-\sigma(-t)$ and $s''(t)=\sigma(t)\sigma(-t)\in\,]0,\frac{1}{4}]$, where $\sigma(t)=1/(1+e^{-t})$. Thus

$$\nabla f(\Theta)=-\frac{1}{L}\sum_\ell\sigma(-c_\ell\Theta^\top u_\ell)c_\ell u_\ell+2\lambda\Theta,\qquad 2\lambda\mathrm{Id}\preceq\nabla^2 f(\Theta)\preceq\Big(\frac{\|U\|^2}{4L}+2\lambda\Big)\mathrm{Id}.$$

f is therefore 2λ-strongly convex and β-smooth with $\beta=\frac{\|U\|^2}{4L}+2\lambda$.

Support vector machines (SVM). Among all hyperplanes $\langle w \,|\, u\rangle + b = 0$ separating the data, we choose the one with maximal margin. The margin boundaries are $\langle w \,|\, u\rangle + b = \pm 1$, so that the width of the margin is $\langle u_+ - u_- \,|\, w/\|w\|\rangle = 2/\|w\|$. Hence

$$\min_{w,b} \ \tfrac{1}{2}\|w\|^2 \quad \text{s.t.} \quad (\forall \ell)\ c_\ell(\langle w \,|\, u_\ell\rangle + b) \geq 1,$$

or, by penalizing constraint violations (*hinge* loss),

$$\min_{w} \ \tfrac{1}{2}\|w\|^2 + \frac{\lambda}{L}\sum_\ell \max\{0, 1 - c_\ell(\langle w \,|\, u_\ell\rangle + b)\}.$$

The hinge loss is convex but **not differentiable** at 1: the problem is strongly convex and nonsmooth (proximal or subgradient algorithms). The same holds for the Lasso $\frac{1}{L}\|U\Theta - c\|^2 + \lambda\|\Theta\|_1$ and sparse logistic regression (ℓ_1 penalty), which favor zero coefficients.

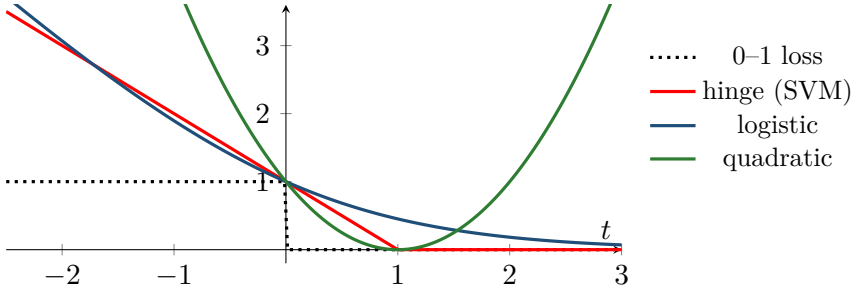


Figure 10: Classification losses as functions of the margin $t = c\ \langle w \,|\, u\rangle$: the logistic loss is smooth (and convex), the hinge loss is convex but not differentiable at $t = 1$; the 0–1 loss is neither convex nor continuous.

6 Exercises

Exercise 1 (Cocoercivity)**.** Show that if f is convex and β-smooth, its gradient is $\frac{1}{\beta}$-cocoercive:

$$\langle \nabla f(x) - \nabla f(y) \,|\, x - y\rangle \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

Exercise 2 (Lipschitz constant)**.** Consider $f(x) = \|Ax - b\|^2 + \|Bx - c\|^2$. Determine the Lipschitz constant of the gradient. What can be deduced about the convergence of the gradient method applied to this function?

Exercise 3 (Tikhonov deconvolution in the Fourier domain)**.** Assume that $\Phi = F^*\Lambda F$ and $D = F^*\Delta F$ are circular convolutions (Λ, Δ diagonal) and consider $f(x) = \frac{1}{2}\|\Phi x - z\|^2 + \frac{\lambda}{2}\|Dx\|^2$.

1. Give the expression of $\widehat{x}$ in the Fourier domain.
2. Express μ and β in terms of Λ and Δ. Under which condition is f strongly convex?
3. Deduce the optimal step size of gradient descent and the associated convergence factor.

References

[1] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, 2004.

[2] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, 2017.

[3] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, 2006.