Robust Optimization & Statistical Learning

Franck Iutzeler

Version: February 11, 2026

Contents

"Predictions are hard, especially about the future"
Uncertainly attributed to Karl Kristian Steincke, Niels Bohr, Yogi Berra, etc.

CHAPTER 1 INTRODUCTION

The purpose of this first part is to clarify our objective. First, we recall the context of decision under uncertainty. Then, we present how robust (linear) optimization problems are connected to statistical learning under distributional uncertainty; this will be our central thread.

Our focus in this monograph is the problem of *decision under uncertainty* and we will turn around the problem of optimizing in the variable[1] $x \in \mathbb{R}^n$ the objective

$$f(x; X) \tag{1.1}$$

where f is a function of suitable input spaces and X is some *uncertain* variable, i.e., a variable that is not perfectly known. In the context of statistical learning, we can think of f as the loss of some machine learning model parametrized by x (e.g. the weights of a neural networks or the coefficients of linear regression model) when facing the data sample X. However, the techniques presented here are not limited nor rooted in the statistical learning community but rather span various domains and applications.

Now, the problem of optimizing the objective (1.1) is ill-defined until we specify how to deal with X. This is a *modeling* issue and several sets of {objectives, assumptions, results, communities} are of independent interest:

- If $X \in \mathsf{X}$ where X is a known set, the *robust optimization* approach to this problem is to solve

$$\min_x \sup_{X \in \mathsf{X}} f(x; X) \tag{RO}$$

  which is usually pessimistic but provides strong guarantees.

- If $X \sim \mu$ where μ is a known probability distribution, the *stochastic programming* approach to this problem is to solve

$$\min_x \mathbb{E}_{X \sim \mu} [f(x; X)] \tag{SP}$$

  which is often more favorable and easier numerically but with looser guarantees.

- If $X \sim \mu$ where $\mu \in \mathsf{U}$ and U is a known set of probability distributions, the *distributionally robust optimization* approach to this problem is to solve

$$\min_x \sup_{\mu \in \mathsf{U}} \mathbb{E}_{X \sim \mu} [f(x; X)] \tag{DRO}$$

  which often appears as performing compromise in statistical learning.

[1] For simplicity, we restrict ourselves to the case of minimization for variables in $\mathbb{R}^n$. Most results can be extended to Hilbert spaces and constrained problems.

We will pay a special attention to the *linear case* i.e., when the objective is a linear functional and the constraint set is formed by linear (in)equalities. Indeed, the problems encountered are already quite rich and there is a lot of intuition to draw from analogy reasoning:

- In (RO), the problem is linear if $f(x; X) = \langle x; X \rangle$ and $\mathcal{X}$ is a linear subset of $\mathbb{R}^n$;
- In (DRO), the *inner* problem is linear as soon as $\mathcal{U}$ is linear. Indeed, the objective is already linear as $\mathbb{E}_{X\sim\mu}[f(x; X)] = \langle f(x; \cdot); \mu \rangle$ where $\langle \cdot; \cdot \rangle$ here denotes the duality pairing between functions and measures.

We will begin by identifying robustness issues in the deterministic and stochastic cases and reviewing classical results about these two approaches. We will also explore how to form distributions sets. Finally, we will explore robustness in statistical learning.

CHAPTER 2 Robust Optimization

Robust optimization has a long history in mathematical programming. Indeed, even for simple linear programs, state-of-the-art solvers can exhibit unstable behaviors in limit cases. We will explore this issue both theoretically and numerically as a means to dive into our topic and explore modern remedies.

Let us first consider the case where the uncertainty is not modeled as a random variable. Problems of the form (RO) are simply constrained optimization problem on the joint variable (x, X). Nevertheless, the study of these problem still has some importance when the size of the perturbation set X is small as it models some uncertainty in the parameters/conditions of the general problem. For instance, in numerical optimization, one can think of rounding errors or approximations in the specification of the objective; in machine learning, having a controlled loss for data samples that are close by can actually make the model more robust.

2.1 Stability in linear programming

Let us consider the linear program

$$\begin{aligned} \inf_{x \in \mathbb{R}^n} \quad & c^\top x \\ \text{subject to} \quad & Gx \leq h \end{aligned} \tag{2.1}$$

where $c \in \mathbb{R}^n$, $G \in \mathbb{R}^{m \times n}$, $h \in \mathbb{R}^m$ and assume that it is feasible i.e., that there exists $x \in \mathbb{R}^n$ such that $Gx \leq h$. Let us call $g_i \in \mathbb{R}^n$ the i-th row of G and $h_i \in \mathbb{R}$ the i-th component of h.

2.1.1 Solutions are on the border of the constraint set

We say that $x^\star$ is a *border point* if $Gx^\star \leq h$ and $\langle g_i, x^\star \rangle = h_i$ for some i.

Then, relying on *optimality conditions*, we have the following result stating that optimal solution are on the border of the polytope defined by the constraints.

Theorem 2.1. *Support that Problem (2.1) is feasible, i.e., that there exists x such that $Gx \leq h$. Then, if $c \neq 0$, exactly one of the following is true:*

(i) the solutions of (2.1) lie on the border (one of the inequalities of the constraint is saturated);

(ii) the infimum of the problem is $-\infty$ (the problem is degenerate).

Proof. The objective and constraint set are convex. Hence, Theorem B.10 states that $x^\star$ is a minimizer if and only if $0 \in c + N_{\{x:Gx\leq h\}}(x^\star)$. This rules out all points in the interior as long as $c \neq 0$.
For a border point $x^\star$, denote $J(x^\star) = \{j : \langle g_j, x^\star\rangle = h_j\}$. Then, $N_{\{x:Gx\leq h\}}(x^\star) = \{\sum_{j\in J(x^\star)} \alpha_j g_j, \alpha \in \mathbb{R}^m_+\}$ (see (Hiriart-Urruty and Lemaréchal, 1993a, Chap. III, Example 5.2.6)).
This means that if $-c$ can be written as a positive combination of the rows $J \subset \{1, .., m\}$ of G, the corresponding solutions are border points, on a face $\{x : \langle g_j, x\rangle = h_j \ \ \forall j \in J\}$.
To prove the converse, we need some intermediate results. □

First, we need a cone separation result. Focusing on (convex) cones provides an elementary geometrical proof for Euclidean spaces which does not invoke Hahn-Banach, we will come to this later on.

We recall that a (linear) cone C in an Euclidean space E is a subset of E that is closed under positive scalar multiplication; that is, $x \in C$ implies $sx \in C$ for every positive scalar s.[2] A cone C is a convex cone if $\alpha x + \beta y$ belongs to C for any positive scalars α, β and any $x, y \in C$. Alternatively, a cone C is convex if and only if $C + C \subseteq C$. Obviously, a convex cone is a convex set.

[2] Usually, it is more convenient to let $0 \in C$, and this is what we will do, but definitions vary in the literature.

Lemma 2.2 (Cone Separation). *Let $C \subset \mathbb{R}^m$ be a nonempty, closed, convex cone and let $b \notin C$. Then there exists a nonzero vector $y \in \mathbb{R}^m$ such that*

$$y^T c \geq 0 \quad \text{for all } c \in C, \qquad \text{and} \qquad y^T b < 0.$$

Proof. Since C is nonempty, closed, and convex, then for any $b \notin C$, there exists a unique point (see Lemma B.3)

$$c_0 = \operatorname{argmin}_{c\in C} \|b - c\|.$$

and since $b \notin C$, we have $c_0 \neq b$.
Let $c \in C$, and define the function

$$\phi(t) = \|b - (c_0 + t(c - c_0))\|^2, \quad t \in [0, 1],$$

which attains a minimum at $t = 0$ since $c_0 + t(c - c_0) = (1 - t)c_0 + tc$ lies in C for $t \in [0, 1]$ by convexity. Differentiating at $t = 0$ on the right yields (see also Theorem B.4)

$$\langle b - c_0,\ c - c_0\rangle \leq 0 \qquad \text{for all } c \in C\,.$$

Now, since C is a cone, $t'c_0 \in C$ for every $t' \geq 0$. Substituting $c = t'c_0$, we obtain

$$\langle b - c_0,\ (t' - 1)c_0\rangle = (t' - 1)\langle b - c_0,\ c_0\rangle \leq 0 \qquad \forall\, t' \geq 0.$$

If $t' > 1$ then $t' - 1 > 0$, giving $\langle b - c_0, c_0\rangle \leq 0$. If $0 < t' < 1$ then $t' - 1 < 0$, giving $\langle b - c_0, c_0\rangle \geq 0$. Thus,

$$\langle b - c_0, c_0\rangle = 0.$$

Hence, we have shown that for any $c \in C$,

$$\langle b - c_0,\ c\rangle = \langle b - c_0,\ c - c_0\rangle + \langle b - c_0,\ c_0\rangle \leq 0\,.$$

Finally, let $y = c_0 - b = -(b - c_0)$. Then,

$$y^T c = -\langle b - c_0,\ c\rangle \geq 0 \qquad \forall\, c \in C,$$

and,

$$y^T b = -\langle b - c_0,\, b\rangle = -\langle b - c_0,\, b - c_0\rangle - \langle b - c_0,\, c_0\rangle = -\|b - c_0\|^2 + 0 < 0\,.$$

Thus y satisfies the claimed inequalities. □

This is essential in geometric proofs of Farkas' lemma which follows. It is often called a lemma of the alternative, as it provides two mutually exclusive and exhaustive conditions.

Lemma 2.3 (Farkas). *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Exactly one of the following two systems has a solution:*

(i) There exists $x \in \mathbb{R}^n$ with $x \geq 0$ such that $Ax = b$.

(ii) There exists $y \in \mathbb{R}^m$ such that $y^T A \geq 0$ (componentwise) and $y^T b < 0$.

Proof. Let $a_1, \ldots, a_n$ denote the columns of A and define the cone

$$C := \{Ax : x \geq 0\} = \left\{\sum_{j=1}^{n} x_j a_j : x_j \geq 0\right\}$$

which is closed and convex.
Suppose (i) holds: $Ax = b$ for some $x \geq 0$. If (ii) also held, then

$$y^T b = y^T Ax = \sum_{j=1}^{n} x_j (y^T a_j) \geq 0,$$

since each $x_j \geq 0$ and each $y^T a_j \geq 0$, contradicting $y^T b < 0$. Thus the systems cannot both be feasible.
Now, assume (i) fails: there is no $x \geq 0$ such that $Ax = b$. Then $b \notin C$. By the cone separation, Lemma 2.2, there exists y such that

$$y^T c \geq 0 \quad \forall c \in C, \qquad \text{and} \qquad y^T b < 0.$$

Since each $a_j \in C$ (take $x = e_j$), we have $y^T a_j \geq 0$ for all j, i.e. $y^T A \geq 0$. Together with $y^T b < 0$, this is exactly system (ii). Thus if (i) fails, (ii) must hold. □

Remark 2.4. Farkas' lemma is the basic theorem of alternative for linear constraints and is fundamental in linear programming duality theory. The geometric content is that b either lies in the cone generated by the columns of A, or there is a hyperplane through the origin separating b from that cone. The Cone Separation Theorem provides exactly the separator needed. ◀

Now, we can proceed with the remainder of the proof.

Proof of Theorem 2.1, continued. Using Farkas' lemma of the alternative (Lemma 2.3; see also (Hiriart-Urruty and Lemaréchal, 1993a, Chap. III, Lemma 4.3.2)), we have that if $-c$ *cannot* be written as a positive combination of the rows $J \subset \{1, .., m\}$ of G, there is $y \in \mathbb{R}^n$ such that $\langle g_j, y\rangle \leq 0$, $j = 1, .., m$ and $\langle c, y\rangle < 0$. Hence, taking $x = \lambda y$ with $\lambda > 0$, we have that $Gx \leq h$ and letting $\lambda \to \infty$, the problem's value tends to $-\infty$. □

Remark 2.5. In Eq. (2.1), we have only inequality constraints. However, nothing prevents taking $g_j = -g_i$ and $h_j = -h_i$ to obtain the equality $\langle g_j, x^\star\rangle = h_j$. ◀

2.1.2 Identification of active constraints with KKT conditions

A more quantitative result for Problem (2.1) can be formulated once we decouple the degenerate inequality constraints. If $g_j = -g_i$ and $h_i = -h_j$, put g_j in matrix C and h_j in vector d and let the remainder of G and h form A and b.

Theorem 2.6. *Support that Problem* (2.1) *is* strictly *feasible, i.e., that there exists x such*

$$Ax < b \text{ and } Cx = d\,.$$

Then, a point x^* is optimal if and only if there exist multipliers $\lambda^* \in \mathbb{R}^m \geq 0$ and $\mu^* \in \mathbb{R}^p$ (where m and p are respectively the size of b and d) such that:

1. $c + A^T\lambda^* + C^T\mu^* = 0$
2. $Ax^* \leq b, \quad Cx^* = d$
3. $\lambda^* \geq 0$
4. $\lambda_i^*(A_i x^* - b_i) = 0$ *for* $i = 1, \ldots, m$.

Proof. This is a direct application of Karush-Kuhn-Tucker conditions (see **??**). □

Remark 2.7 (Link with Theorem 2.1). The fourth condition identifies inequalities that are saturated (if any) : those for which $\lambda_i^* > 0$. ◂

2.1.3 Solutions are almost always extremal points

For Lebesgue almost all c, solutions are extremal points, so changes in constraints lead to drastic changes in solutions. To see that, let

$$P \;=\; \{x \in \mathbb{R}^n : Gx \leq h\}\,,$$

be the *polyhedron* of constraints.[3]

[3] If it is bounded, it is called a *polytope*.

Definition 2.8 (Extreme point)**.** Let $C \subset \mathbb{R}^n$ be a convex set. A point $v \in C$ is an *extreme point* of C if

$$v = \lambda x + (1-\lambda)y, \quad x, y \in C,\ 0 < \lambda < 1 \quad \Longrightarrow \quad x = y = v.$$

Equivalently, v cannot be written as a nontrivial convex combination of two distinct points of C.

In our case, a point $v \in P$ is a *vertex* (extreme point) of P if it cannot be expressed as a convex combination of two distinct points in P. Denote by $V = \{v^1, \ldots, v^K\}$ the (finite) set of all vertices of P (see **????**).

Theorem 2.9 (Minkowski, 1896; Convex hull representation)**.** *Let $P \subset \mathbb{R}^n$ be a nonempty polytope (bounded polyhedron), and let V denote the set of its extreme points. Then*

$$P = \mathrm{conv}(V),$$

i.e., every point in P is a convex combination of its vertices.

Remark 2.10. This theorem is fundamental in linear programming: for any linear functional $c^\top x$,

$$\max_{x \in P} c^\top x = \max_{v \in V} c^\top v,$$

so the optimum is attained at a vertex. ◂

Theorem 2.11 (Generic optimality at vertices)**.** *With the notation above, the set*

$$S := \{c \in \mathbb{R}^n : LP(c) \text{ has no optimal solution that is a vertex of } P\}$$

is contained in the finite union of proper linear hyperplanes

$$\bigcup_{1 \le i < j \le K} \{c : c^\top (v^i - v^j) = 0\}.$$

Consequently S has Lebesgue measure zero, and for almost every $c \in \mathbb{R}^n$ the program $LP(c)$ has a unique optimal solution and that solution is a vertex of P.

Proof. Because P is a bounded polyhedron, every point $x \in P$ is a convex combination of vertices of P (Minkowski's theorem). In particular, whenever an optimal solution exists, there is an optimal solution that is a convex combination of optimal vertices; equivalently the maximum of the linear functional $x \mapsto c^\top x$ over the compact set P is attained at (at least) one vertex.
Thus it suffices to characterize those c for which *no* vertex attains the maximum (equivalently, every optimal solution lies strictly in a face of dimension ≥ 1 and at least two vertices tie).
Fix two distinct vertices $v^i \ne v^j$. The condition that v^i and v^j produce the same objective value is

$$c^\top v^i = c^\top v^j \quad \Longleftrightarrow \quad c^\top (v^i - v^j) = 0.$$

The solution set of this linear equation is a proper linear hyperplane in $\mathbb{R}^n$ (proper because $v^i - v^j \ne 0$).
If $\mathrm{LP}(c)$ has two distinct optimal vertices v^i and v^j, then necessarily c satisfies $c^\top (v^i - v^j) = 0$. Hence the set of c for which there are at least two optimal vertices is contained in the finite union

$$\bigcup_{1 \le i < j \le K} \{c : c^\top (v^i - v^j) = 0\}.$$

Therefore, if c lies outside this union, the objective values $\{c^\top v^i\}_{i=1}^K$ are all distinct and the maximum over the finite set V is achieved at a unique vertex $v^{i(c)}$. Because the maximum over P equals the maximum over V, it follows that $\mathrm{LP}(c)$ has a unique optimal solution and this solution is the vertex $v^{i(c)}$.
Finally, a finite union of proper hyperplanes has Lebesgue measure zero in $\mathbb{R}^n$. Hence the exceptional set S (which is contained in that finite union) has measure zero, proving the claim. □

Remark 2.12. Uniqueness of the optimal solution implies that the solution is a vertex: if the unique maximizer x^* were not extreme then $x^* = \frac{1}{2}(x_1 + x_2)$ for some distinct $x_1, x_2 \in P$, and by linearity of the objective both x_1 and x_2 would also be optimal, contradicting uniqueness. ◂

Remark 2.13. For general (possibly unbounded) polyhedra, the set of vertices is still finite. Boundedness is not required for finiteness; it is required only to guarantee that P is nonempty and that the LP optimum is attained. The finiteness of vertices follows solely from the fact that any vertex must be determined by n linearly independent tight inequalities. ◂

2.1.4 Theory versus practice

Now, let us identify a stability problem by considering the problem

$$\begin{aligned} \inf_{x\in\mathbb{R}^2} \quad & x_1 + x_2 \\ \text{subject to} \quad & (1+X)x_1 + x_2 \geq 1 \\ & x_1 + (1-X)x_2 \geq 1 \\ & x_1 + x_2 = 1 \\ & x_1, x_2 \geq 0 \end{aligned} \tag{2.2}$$

for some $X \in [-0.5, 0.5]$.

It is easy to see that $[1, 0]$ is a solution if $X \geq 0$ and that $[0, 1]$ is a solution if $X \leq 0$. This means that if X is uncertain, the solutions of the problem can change. Nevertheless, here the value of the problem (i.e., the value of $x_1 + x_2$) *does not* change with δ, only the *chosen solution* changes.

Though the problem's value does not change, this abrupt change of optimal point is an issue both numerically (see after) and in practice (as it can lead to opposite decisions).

Example 2.14 (Numerical solutions). Using Scipy's linprog solver, the problem above is solved by the following code.

```
from scipy.optimize import linprog

delta =  -1e-7

c = [1,1]

A_ub = [[-1*(1+delta),-1],[-1,-1*(1-delta)]]
b_ub = [-1,-1]

A_eq = [[1,1]]
b_eq = [1]

l = 0
u = None

res = linprog(c, A_ub=A_ub, b_ub=b_ub, A_eq=A_eq, b_eq = b_eq,
    bounds=(l,u))

print(res.message)
print(res.x)
```

With $\delta = -1e-7$, the obtained solution is $[1, 0]$ which is unfeasible ! As soon, as I take $\delta = -1.1e-7$, the solution jumps to $[-0, 1]$...

2.1.5 Conclusion on robustness in linear programming

- Solutions of linear programs belong to the border of the constraint set, they are even almost always extremal points.
- A change in the active contraints will directly lead to a change in the solution. If a constraint is removed, the change can be very brutal.
- A change in the objective may or may not lead to a change in the solution. If c belong to the interior of the normal cone to the constraints at the solution, the

solution stays unchanged for any closeby c, which is a form of robustness.

Remark 2.15 (Robust (Linear) Optimization). Some explicit form of robustness can be added directly in the formulation of the problem, we then talk about robust linear optimization. Two good references fon the topic are (Bertsimas et al., 2011) and (Ben-Tal et al., 2009). The idea is to reformulate the optimization problem in order to explicitly model the uncertainties of the problem and directly take care of them. A good practical insight can be found in the documentation of the solver Mosek. This adds complexity to the problem to solve both numerically and theoretically. Indeed, in Problem (2.2), the solution $[0.5, 0.5]$ appears to be a good compromise *but* it is unfeasible for any $X \neq 0$! ◀

2.2 Smoothing & Differentiable programming

Differentiable programming is a cornerstone of modern robust statistical learning, providing the computational tools necessary to tackle complex, uncertain, and data-driven problems efficiently. It is rooted in the ability to compute gradients for all operations, even when differentiability is not present.

This is done by extending traditional non-differentiable operations such as maximums, logical connectors, flooring, etc.

2.2.1 Differentiable extensions of logical operations

Replacing a boolean $\Pi \in \{0, 1\}$ by a continuous variable $\pi \in [0, 1]$ representing the "probability of being true" is rather common in statistical learning (Furthermore, to extended the values to $\mathbb{R}$, one can take $\pi = \text{sigmoid}(q) := 1/(1 + \exp(-q))$).

Similarly, logical operators can be extended in the same spirit:

$$\begin{aligned} \text{and}(\pi, \pi') &:= \pi \cdot \pi' \\ \text{or}(\pi, \pi') &:= \pi + \pi' - \pi \cdot \pi' \\ \text{not}(\pi) &:= 1 - \pi \\ \text{ifthenelse}(\pi, v_{\text{true}}, v_{\text{false}}) &:= \pi \cdot v_{\text{true}} + (1 - \pi) \cdot v_{\text{false}} \end{aligned}$$

2.2.2 Optimization of a supremum & differentiability

If the set $\mathcal{X}$ in which X lives is compact, we can obtain the differentiability of the supremum by computing the gradient of f with respect to x at the point attaining the maximum.

More precisely, from $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, we define the functions

$$h(x) := \sup_{X \in \mathcal{X}} f(x; X) \quad \text{and} \quad X^\star(x) := \text{argmax}_{X \in \mathcal{X} \subset \mathbb{R}^p} f(x; X).$$

In order to study the derivatives of maximums, the two following results leads to the the same conclusion with two different sets of assumptions.

Theorem 2.16 (Rockafellar's envelope theorem). *Let $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be jointly C^1 and let $\mathcal{X}$ be a compact convex subset of $\mathbb{R}^p$. Then, if $X^\star(x)$ is unique, then h is differentiable and*

$$\nabla h(x) = \nabla_x f(x; X^\star(x)) .$$

[4] *i.e., concave in x convex in y*

Theorem 2.17 (Danskin's theorem). *Let $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be concave-convex[4] and let X be a compact convex subset of $\mathbb{R}^p$. Then, if $X^\star(x)$ is unique, then h is differentiable and*

$$\nabla h(x) = \nabla_x f(x; X^\star(x)) .$$

Example 2.18 (Convex conjugation). Let $f(x, y) = \langle x, y\rangle - \Omega(x)$ and $\mathcal{X} = \Delta_n$. Then $h(y) = \Omega^*(y)$ where Ω^* denotes the convex conjugate of Ω. Then, the $x^\star(y)$ is unique as soon as Ω is strictly convex (for instance) (Hiriart-Urruty and Lemaréchal, 1993a, Th X.4.1.1). Then, $\nabla h(y) = x^\star(y) = \nabla\Omega^*(y)$. Furthermore, if Ω is μ strongly convex, then $\nabla\Omega^*$ is $1/\mu$-Lipschitz continuous (Hiriart-Urruty and Lemaréchal, 1993a, Th X.4.2.2).

This is heavily used in game theory and adversarial learning which will talk about later on. However, when this is not the case, one has to resort to smoothing.

2.2.3 Smoothing of the maximum and argmax

Given a vector $u \in \mathbb{R}^n$, we define its maximum/max and argmax as

$$\max(u) := \max_{j\in\{1,..,n\}} u_j \quad \text{and} \quad \operatorname{argmax}(u) := \{i : u_i = \max_{j\in\{1,..,n\}} u_j\}$$

where the max is real valued while the argmax has between one and n outputs in $\{1, .., n\}$.[5]

[5] Note that one could define $\operatorname{argmax}(u) := \{u_i : u_i = \max_{j\in\{1,..,n\}} u_j\}$ i.e., to output the entry rather than the coordinate, but the entry goes better with smoothing.

As classical in smoothing (see the logical operators above), it is natural to replace a choice of alternative (e.g. a coordinate) with a probability distribution on the alternatives. We recall the notation $\Delta_n := \{\pi \in \mathbb{R}^n : \pi \geq 0, \sum_{i=1}^n \pi_i\}$ for the n-simplex, i.e., the set of probability distributions on n elements.

Lemma 2.19. *We have*

$$\max(u) = \max_{\pi\in\Delta_n} \langle u, \pi\rangle = \max_{\pi\in\{e_1,..,e_n\}} \langle u, \pi\rangle$$

and

$$\operatorname{argmax}(u) = \operatorname{argmax}_{\pi\in\Delta_n} \langle u, \pi\rangle = \operatorname{argmax}_{\pi\in\{e_1,..,e_n\}} \langle u, \pi\rangle$$

Proof. See **??**. □

The rationale in differentiable programming and smoothing is that all components of the probability vector should have a positive probability. In some sense, we are adding noise to the output of the max operator. In order to do so in a controlled manner, we have to mitigate the objective vs noise.

Lemma 2.20. *The entropy-regularized maximum operator, also called softmax, is the log-sum-exp mapping. For $\lambda > 0$, we have*

$$\begin{aligned}
\operatorname{soft}\max(u) &= \max_{\pi\in\Delta_n} \langle u, \pi\rangle + \lambda H(\pi) \\
&= \max_{\pi\in\Delta_n} \langle u, \pi\rangle - \lambda \sum_{i=1}^n \pi_i \log(\pi_i) \\
&= \lambda \log \sum_{i=1}^n \exp(u_i/\lambda)
\end{aligned}$$

Proof.

$$\begin{aligned}
\text{soft max}(u) &= \max_{\pi \in \Delta_n} \langle u, \pi \rangle - \lambda \sum_{i=1}^n \pi_i \log(\pi_i) \\
&= \max_{\pi \in \mathbb{R}^n_+} \min_{t \in \mathbb{R}} \langle u, \pi \rangle - \lambda \sum_{i=1}^n \pi_i \log(\pi_i) - t(\sum_{i=1}^n \pi_i - 1) \\
&= \min_{t \in \mathbb{R}} \max_{\pi \in \mathbb{R}^n_+} \langle u, \pi \rangle - \lambda \sum_{i=1}^n \pi_i \log(\pi_i) - t(\sum_{i=1}^n \pi_i - 1) \\
&= \min_{t \in \mathbb{R}} t + \max_{\pi \in \mathbb{R}^n_+} \sum_{i=1}^n \pi_i (u_i - \lambda \log(\pi_i) - t) \\
&= \min_{t \in \mathbb{R}} t + \sum_{i=1}^n \max_{p \in \mathbb{R}_+} p(u_i - \lambda \log(p) - t)
\end{aligned}$$

where the first two equalities comes from the Lagrange Duality. Now, maximizing $p(u_i - \lambda \log(p) - t)$ in p leads to $p_i^\star = \exp((u_i - t - \lambda)/\lambda)$. We are left with

$$\text{soft max}(u) = \min_{t \in \mathbb{R}} t + \sum_{i=1}^n \exp((u_i - t - \lambda)/\lambda)\lambda$$

and nulling the gradient of the objective (in t) leads to the equation

$$\begin{aligned}
1 &= \sum_{i=1}^n \exp((u_i - t - \lambda)/\lambda) \\
&= \exp(-t/\lambda) \exp(-1) \sum_{i=1}^n \exp(u_i/\lambda)
\end{aligned}$$

and thus $t^\star = \lambda \log\left(\sum_{i=1}^n \exp(u_i/\lambda)\right) - \lambda$. Plugging into problem, we obtain the claimed result. □

Now, we have a differentiable approximation of the maximum and thus can link the gradient of the maximum to the softargmax. Furthermore, we can show smoothness of the problem.

Proposition 2.21. *For $\lambda > 0$, we have*

$$\begin{aligned}
\nabla \text{soft max}(u) = \text{soft argmax}(u) &= \text{argmax}_{\pi \in \Delta_n} \langle u, \pi \rangle + \lambda H(\pi) \\
&= \left[\frac{\exp(u_j/\lambda)}{\sum_{i=1}^n \exp(u_i/\lambda)} \right]_j
\end{aligned}$$

and ∇soft max is $1/\lambda$-Lipschitz continuous

Proof. For the first part, take the value of $p_i^\star$ on the proof of the previous lemma with the right value for $t^\star$. For the second part, compute the Hessian. □

Exercises

Exercise 2.1. Show that that the problem $\max_{\pi \in \Delta_n} \langle u, \pi \rangle + \lambda H(\pi)$ is equivalent to the problem $\max_{\pi \in \Delta_n : H(\pi) \geq \varepsilon} \langle u, \pi \rangle$ i.e., that for any ε there is λ so that the problems have the same solution and vice versa.

Exercise 2.2. Show that that $\max(u) \leq \text{soft}\max(u) \leq \max(u) \log(n)$ using properties of the entropy.

Exercise 2.3. Show that that $\text{soft}\max([\text{soft}\max([a,b]), c]) = \text{soft}\max([a, \text{soft}\max([b,c])])$.

Exercise 2.4. Entropy is not the only relaxation possible. Using Gini's negentropy $\Omega(\pi) := 1/2 \sum_{i=1}^n \pi_i(\pi_i - 1)$, show that we obtain the sparsemax operator

$$\text{sparse}\max(u) = \min_t t + \frac{1}{2} \sum_{i=1}^n [u_i - t]_+^2$$

Exercise 2.5. In Example 2.18, the linear functional $\langle x, y \rangle$ is optimized over the compact set Δ_n. What is a good choice for Ω to make the max and argmax of $f_y(x) = \langle x, y \rangle - \Omega(x)$ differentiable?

CHAPTER 3 Game Theory

Game theory is a set of analytical tools to understand the phenomena observed when decision-makers interact. Interestingly, this will enable us to see how to optimize a function depending on unknown external parameters (the players' actions) when all players *behave rationally.*

The *players* pursue well-defined objectives and take into account what they know of the other players' behavior. A *game* is the description of the players, their possible actions, and their interest. The modelling/formalization is very important.

A bit of history:

- Traces since 1713 by Waldegrave, for the analysis of a card game;
- Renewed interest in the 1920s with chess analysis;
- Von Neumann's "On the theory of Games of Strategy"(von Neumann, 1928) in 1928 kickstarted the field;
- Nobel prizes (economy mostly) in 1994 (inc. John Nash), 2005, 2007, 2012, and 2015 (Jean Tirole).

This part is mainly based on (Osborne and Rubinstein, 1994).

3.1 Games & Strategies

Games are commonly classified according to the number of players involved, the nature of their strategy spaces, and the structure of interactions among players.

Game Type	Players and Strategy Spaces	Interaction Structure	Typical Applications
Finite Games	Finite number of players; each player has a finite set of strategies	Direct strategic interaction among all players; payoffs depend on full strategy profiles	Classical economics, auctions, bargaining, voting models
Population Games	Large (often infinite) population of players; strategies chosen from a finite set	Anonymous interaction; payoffs depend on aggregate population behavior rather than individual identities	Evolutionary biology, traffic routing, social dynamics
Continuous Games	Finite number of players; strategy sets are continuous (e.g. intervals in $\mathbb{R}^n$)	Payoffs are continuous functions of players' strategy choices	Oligopoly models, pricing and quantity competition, resource allocation
Mean Field Games	Continuum (or very large number) of players; continuous state and control spaces	Each player interacts with the distribution (mean field) of states/actions in the population	Economics with many agents, finance, crowd dynamics, energy systems

These classes of games provide a spectrum of modeling frameworks, ranging from small-scale strategic interactions to large-population limits. The choice of model depends on the scale of the system, the desired level of analytical tractability, and the nature of the interactions being studied.

We will mainly consider finite games in this chapter.

3.2 Finite Games

3.2.1 Notation

There is a finite set of *players* $\mathrm{P} = \{1, .., N\}$. Each player i has a *set of actions* S_i and a *payoff function* $\mathrm{g}_i : \mathrm{S}_1 \times .. \times \mathrm{S}_N \to \mathbb{R}$.

Definition 3.1. A game in *normal form* is a tuple $\Gamma = (N, \mathrm{S} = \{\mathrm{S}_i\}, \mathrm{g} = \{\mathrm{g}_i\})$.

In a *pure strategy*, each player i chooses *one* action $\mathrm{s}_i \in \mathrm{S}_i$. Then, it receives the payoff $\mathrm{g}_i(\mathrm{s}_1, .., \mathrm{s}_N)$.

We will also note:
$\mathrm{S} = \mathrm{S}_1 \times \mathrm{S}_2 \times .. \times \mathrm{S}_N$
$\mathrm{S}_{-i} = \bigtimes_{j \neq i} \mathrm{S}_j$
$\mathrm{g} = (\mathrm{g}_i)_i$

3.2.2 Different types of games

We will illustrate several types of fundamental games that capture the diversity of normal games. Each time, we will exhibit a two players game ($N = 2$) as they can easily be represented graphically and are the most basic and insightful examples in game theory.

They are typically represented as a table:

		Player 2	
		s_2	s'_2
Player 1	s_1	$(g_1(s_1, s_2), g_2(s_1, s_2))$	$(g_1(s_1, s'_2), g_2(s_1, s'_2))$
	s'_1	$(g_1(s'_1, s_2), g_2(s'_1, s_2))$	$(g_1(s'_1, s'_2), g_2(s'_1, s'_2))$

Common interest A game where the players have the same payoff: $g_i = g_j$ for all $i, j \in P$.

Example 3.2 (Activity in Grenoble). Alice and Bob want to do something together, either trail T or ski S with no preference.

$S_A = S_B = \{T, S\}$ and $g_A = g_B = \begin{cases} 1 & \text{if } s_A = s_B \\ 0 & \text{else} \end{cases}$

Zero-sum games A game where the player are antagonist: $\sum_{i=1}^{N} g_i \equiv 0$

Example 3.3 (Matching pennies). Alice and Bob both have a penny; they secretly turn it to heads or tails. If the pennies match, Alice wins 1€ and Bob loses 1€ (Bob gives 1€ to Alice). If they are different Alice gives 1€ to Bob.

$S_A = S_B = \{H, T\}$ and $g_A = -g_B = \begin{cases} 1 & \text{if } s_A = s_B \\ -1 & \text{else} \end{cases}$

Battle of the sexes Mix between common interest and zero-sum.

Example 3.4 (Meetup). Alice and Bob want to meet tonight; Alice prefers to meet at a bar; Bob prefers to meet at home.

$S_A = S_B = \{B, H\}$, $g_A = \begin{cases} 3 & \text{if } s_A = s_B = B \\ 1 & \text{if } s_A = s_B = H \\ 0 & \text{else} \end{cases}$, $g_B = \begin{cases} 1 & \text{if } s_A = s_B = B \\ 3 & \text{if } s_A = s_B = H \\ 0 & \text{else} \end{cases}$

Prisonner's dilemma It is a classic game where Alice and Bob are arrested and individually given the possibility to stay silent or cooperate.

$S_A = S_B = \{S, C\}$,

$$g_A = \begin{cases} -1 & \text{if } s_A = S \text{ and } s_B = S \\ -3 & \text{if } s_A = S \text{ and } s_B = C \\ 0 & \text{if } s_A = C \text{ and } s_B = S \\ -2 & \text{if } s_A = C \text{ and } s_B = C \end{cases}$$

$$g_B = \begin{cases} -1 & \text{if } s_A = S \text{ and } s_B = S \\ 0 & \text{if } s_A = S \text{ and } s_B = C \\ -3 & \text{if } s_A = C \text{ and } s_B = S \\ -2 & \text{if } s_A = C \text{ and } s_B = C \end{cases}$$

It is a fundamental game in economy, notably for the creation of rules enabling the denunciation of coalitions between companies.

Game of chicken A lot like the prisonner's dilemma but penalizing a lot mutual cooperation.

$S_A = S_B = \{S, C\}$,

$$g_A = \begin{cases} -1 & \text{if } s_A = S \text{ and } s_B = S \\ -3 & \text{if } s_A = S \text{ and } s_B = C \\ 0 & \text{if } s_A = C \text{ and } s_B = S \\ -20 & \text{if } s_A = C \text{ and } s_B = C \end{cases}$$

$$g_B = \begin{cases} -1 & \text{if } s_A = S \text{ and } s_B = S \\ 0 & \text{if } s_A = S \text{ and } s_B = C \\ -3 & \text{if } s_A = C \text{ and } s_B = S \\ -20 & \text{if } s_A = C \text{ and } s_B = C \end{cases}$$

It is the game modeling mutually assured destruction: cuban missile crisis, evolutionary biology, etc.

3.2.3 Dominated pure strategies

Definition 3.5. A strategy $s_i \in S_i$ is *dominated* if there is $t_i \in S_i$ such that

$$\forall s_{-i} \in S_{-i}, \quad g_i(t_i; s_{-i}) \geq g_i(s_i; s_{-i}).$$

It is *strictly dominated* if the inequality is strict.

A rational player never plays a strictly dominated strategy.

Definition 3.6. A strategy $s_i \in S_i$ is *dominating* if for all $t_i \in S_i$

$$\forall s_{-i} \in S_{-i}, \quad g_i(s_i; s_{-i}) \geq g_i(t_i; s_{-i}).$$

It is *strictly dominating* if the inequality is strict.

It is unique from definition. If it exists, it is the only rational action.

Example 3.7. What should player 1 play in the following game?

		Player 2	
		A	B
Player 1	A	$(0, -2)$	$(-10, -1)$
	B	$(-1, -10)$	$(-5, -5)$

- What will play Player 2?
- Deduce what should play Player 1.
- Is it the best payment both player could have had?

If there exists *dominated strategies*, they can be eliminated successively from the game.

3.2.4 Nash Equilibrium

Definition 3.8. A strategy profile $s = s_1 \times s_2 \times .. \times s_N \in S$ is a *Nash Equilibrium* (NE) if

$$\forall i, \forall t_i \in S_i, \quad g_i(s_i; s_{-i}) \geq g_i(t_i; s_{-i}).$$

It is a global equilibrium (contrary to the local ones seen before). No player has a singular interest to deviate from his action. It is thus a good way to conclude an agreement.

Are there Nash equilibriums in the following games?

Common Interest

		Bob	
		T	S
Alice	T	$(1,1)$	$(0,0)$
	S	$(0,0)$	$(1,1)$

Zero Sum

		Bob	
		H	T
Alice	H	$(1,-1)$	$(-1,1)$
	T	$(-1,1)$	$(1,-1)$

Battle of the sexes

		Bob	
		B	H
Alice	B	$(3,2)$	$(0,0)$
	H	$(0,0)$	$(2,3)$

Prisonner's dilemma

		Bob	
		Silent	Cooperate
Alice	Silent	$(-1,-1)$	$(-3,0)$
	Cooperate	$(0,-3)$	$(-2,-2)$

Game of Chicken

		Bob	
		Silent	Cooperate
Alice	Silent	$(-1,-1)$	$(-3,0)$
	Cooperate	$(0,-3)$	$(-20,-20)$

Remark 3.9 (Nash Equilibriums and dominating strategies).

- There can be no, one, or several NEs.
- If there is a strictly dominating strategy matching each player, it is the unique NE.
- By eliminating successively strictly dominated strategies, NEs are preserved.
- A profile of dominating strategies is a NE.

◀

Remark 3.10 (Continuous games). These concepts can be directly extended to continuous actions sets, see Exercices 3.3 and 3.4. ◀

3.2.5 Equilibrium Selection

a)

		Player 2	
		A	B
Player 1	A	$(9, 9)$	$(-15, 8)$
	B	$(8, -15)$	$(7, 7)$

(A,A) and (B,B) are two NEs. If the player are risk-averse, they may prefer (B,B) even though the payoff is smaller. Indeed, if the other player does not play the NE, the loss is smaller with (B,B).

b)

		Player 2	
		A	B
Player 1	A	$(2, 2)$	$(1, 1)$
	B	$(1, 1)$	$(1, 1)$

(A,A) and (B,B) are two NEs but B is dominated for each player while A is strictly dominating. So (A,A) seems better.

c)

		Player 2	
		A	B
Player 1	A	$(2, 2)$	$(1, 2)$
	B	$(2, 1)$	$(1, 1)$

All states are NEs!

3.3 Mixed strategies

For some games, NEs *with pure strategies* do not exist; for instance, in Rock-Paper-Scissors.

Example 3.11 (Rock-Paper-Scissors).

		Player 2		
		Rock	Paper	Scissors
Player 1	Rock	$(0, 0)$	$(-1, 1)$	$(1, -1)$
	Paper	$(1, -1)$	$(0, 0)$	$(-1, 1)$
	Scissors	$(-1, 1)$	$(1, -1)$	$(0, 0)$

3.3.1 Mixed games

Let $\Gamma = (N, S = \{S_i\}, g = \{g_i\})$ be a game in normal form and let us suppose that *each* S_i *is a finite set.*

Definition 3.12. A *mixed strategy* σ_i for player i is a probability distribution on S_i.

$$\sigma_i = (\sigma_i(S_i[1]), .., \sigma_i(S_i[n_i])) \in \Delta(S_i)$$

where $\sigma_i(S_i[j]) = \mathbb{P}[i \text{ plays the } j\text{-th action in his set}]$ and $\Delta(S_i)$ is the simplex[6] on S_i.

[6] The Simplex Δ_n of size n is the set of all vector of $\mathbb{R}^n$ such that $x_i \geq 0$ and $\sum_{i=1}^n x_i = 1$.

Interpretation:

- Random strategy (eg in Rock Paper Scissors)
- Model for a large number of players

We note $\Sigma = \bigtimes_i \Delta(S_i)$ and $\Sigma_{-i} = \bigtimes_{j \neq i} \Delta(S_j)$.

Mixed game

- Each player plays a mixed strategy $\sigma_i \in \Delta(S_i)$.
- The probability that the global strategy $s = (s_1, .., s_N)$ is played is $\prod_j \sigma_j(s_j)$.
- For a global strategy $\sigma \in \Sigma$, the *expected payoff* for player i is

$$g_i(\sigma) = \mathbb{E}_{s \sim \Sigma}[g_i(s)] = \sum_{s \in S} \left[\prod_j \sigma_j(s_j) \right] g_i(s).$$

With these definitions, $\Gamma = (N, \Sigma = \{\sigma_i\}, g = \{g_i\})$ is a *mixed game*:

- The players simultaneously choose a pure strategy $s_i \sim \sigma_i$
- They get payoff $g_i(s)$
- Each player tries to maximize its expected payoff

3.3.2 Nash Equilibriums for Mixed Games

Definition 3.13. A mixed strategy profile $\sigma = \sigma_1 \times \sigma_2 \times .. \times \sigma_N \in \Sigma$ is a *Nash Equilibrium* (NE) if

$$\forall i, \forall \tau_i \in \Sigma_i = \Delta(S_i), \quad g_i(\sigma_i; \sigma_{-i}) \geq g_i(\tau_i; \sigma_{-i}).$$

Example 3.14 (Rock-Paper-Scissors). (1/3,1/3,1/3) is a NE.

Theorem 3.15 (Nash's Theorem (1950)). *All finite[7] games have (mixed) Nash Equilibriums.*

[7] *with finite number of actions*

Proof. Upcoming! □

3.3.3 Dominated Mixed Strategies

Definition 3.16. A mixed strategy $\sigma_i \in \Sigma_i$ is *dominated* if there is $\tau_i \in \Sigma_i = \Delta(S_i)$ such that

$$\forall \sigma_{-i} \in \Sigma_{-i}, \quad g_i(\tau_i; \sigma_{-i}) \geq g_i(\sigma_i; \sigma_{-i}).$$

It is *strictly dominated* if the inequality is strict.

Example 3.17.

		Player 2		
		A	B	C
Player 1	a	$(1,1)$	$(0,2)$	$(0,4)$
	b	$(0,2)$	$(5,0)$	$(1,6)$
	c	$(0,2)$	$(1,1)$	$(2,1)$

Question 3.1. Show that for Player 2, strategy B is strictly dominated by $0.5A + 0.5C$.

While we could remove strictly dominated mixed strategy, this does not lead to a reduction of the states of the game. However, we are still able to remove strictly dominated *pure* strategies.

Proposition 3.18. *Let (Γ^k) be the sequence of games produced by eliminating strictly dominated pure strategies in Γ. Then, for all k, $NE(\Gamma^k) = NE(\Gamma)$.*

Example 3.19. We saw in Example 3.17 that B was strictly dominated by mixed strategy $0.5A + 0.5C$, thus we can remove it

		Player 2	
		A	C
Player 1	a	$(1,1)$	$(0,4)$
	b	$(0,2)$	$(1,6)$
	c	$(0,2)$	$(2,1)$

We can remove b for player 1 since it is dominated by eg. $0.4a + 0.6b$. We obtained a reduced game:

		Player 2	
		A	C
Player 1	a	$(1,1)$	$(0,4)$
	c	$(0,2)$	$(2,1)$

The Nash Equilibrium of the original game is $(1/4, 0, 3/4)$ for Player 1 and $(2/3, 0, 1/3)$ for Player 2. We will see how to find it in the forthcoming sections.

3.3.4 Looking for mixed equilibriums

Definition 3.20. For player i, $\sigma_i \in \Sigma_i$ is a *best response* to $\sigma_{-i} \in \Sigma_{-i}$ if

$$\forall \tau_i \in \Sigma_i = \Delta(S_i), \quad g_i(\sigma_i; \sigma_{-i}) \geq g_i(\tau_i; \sigma_{-i}).$$

The set of all best responses for an adversarial strategy $\sigma_{-i} \in \Sigma_{-i}$ is denoted by $\mathrm{BR}(\sigma_{-i})$

The following result is obvious from the definitions.

Proposition 3.21. *$\sigma \in \Sigma$ is a (mixed) Nash Equilibrium if and only if for all i, $\sigma_i \in \mathrm{BR}(\sigma_{-i})$.*

There is a nice relation between pure and mixed strategies in terms of best response. To study it, let use denote the *support* of a mixed strategy as $\mathrm{supp}(\sigma_i) = \{s_i \in S_i : \sigma_i(s_i) > 0\}$, i.e. the actions that have a positive probability to be played.

Proposition 3.22 (Weak Indifference). *For player i, an adversarial strategy $\sigma_{-i} \in \Sigma_{-i}$, and $\sigma_i \in \mathrm{BR}(\sigma_{-i})$, then*

$$\forall s_i \in \mathrm{supp}(\sigma_i), \quad g_i(s_i; \sigma_{-i}) = g_i(\sigma_i; \sigma_{-i}).$$

This mean that all pure strategies in support have the same payoff, equal to the payoff of the mixed strategy.

Proof.

$$g_i(\sigma_i; \sigma_{-i}) = \sum_{s_i \in S_i} \sigma_i(s_i) g_i(s_i; \sigma_{-i}) = \sum_{s_i \in \mathrm{supp}(\sigma_i)} \sigma_i(s_i) g_i(s_i; \sigma_{-i})$$

Then:

1) $g_i(s_i; \sigma_{-i}) \leq g_i(\sigma_i; \sigma_{-i})$ since $\sigma_i \in \mathrm{BR}(\sigma_{-i})$;
2) Suppose that there is $t_i \in \mathrm{supp}(\sigma_i)$ such that $g_i(t_i; \sigma_{-i}) < g_i(\sigma_i; \sigma_{-i})$. Then,

$$\begin{aligned} g_i(\sigma_i; \sigma_{-i}) &= \sum_{s_i \in S_i} \sigma_i(s_i) g_i(s_i; \sigma_{-i}) \\ &< \sum_{s_i \in \mathrm{supp}(\sigma_i)} \sigma_i(s_i) g_i(\sigma_i; \sigma_{-i}) \text{ (by our supposition)} \\ &= g_i(\sigma_i; \sigma_{-i}) \text{ (since } \sigma_i \text{ is a probability vector)} \end{aligned}$$

which is a absurd.

Hence, $g_i(s_i; \sigma_{-i}) = g_i(\sigma_i; \sigma_{-i})$ for all $s_i \in \mathrm{supp}(\sigma_i)$. □

The notion of indifference can be strengthened as follows.

Proposition 3.23 (Strong Indifference). *For player i and an adversarial strategy $\sigma_{-i} \in \Sigma_{-i}$,*

$$\sigma_i \in \mathrm{BR}(\sigma_{-i}) \iff \begin{cases} (1) & \forall s_i, t_i \in \mathrm{supp}(\sigma_i), \quad g_i(s_i; \sigma_{-i}) = g_i(t_i; \sigma_{-i}) \\ (2) & \forall s_i \notin \mathrm{supp}(\sigma_i), \quad g_i(s_i; \sigma_{-i}) \leq g_i(\sigma_i; \sigma_{-i}) \end{cases}.$$

Proof. The forward way is direct from the previous proof. The other way comes from noticing that (1) + (2) imply that $g_i(s_i; \sigma_{-i}) \leq g_i(\sigma_i; \sigma_{-i})$ for all $s_i \in S_i$ and thus σ_i is a best response to σ_{-i}. □

Using once again the link between best responses and Nash Equilibriums, we have the following result.

Corollary 3.24. *The strategy $\sigma \in \Sigma$ is a (mixed) Nash Equilibrium if and only if for each player i:*

$$\begin{cases} (1) & \forall s_i, t_i \in \mathrm{supp}(\sigma_i), \quad g_i(s_i; \sigma_{-i}) = g_i(t_i; \sigma_{-i}) \\ (2) & \forall s_i \notin \mathrm{supp}(\sigma_i), \quad g_i(s_i; \sigma_{-i}) \leq g_i(\sigma_i; \sigma_{-i}) \end{cases}.$$

Thus, in order to find Nash Equilibriums:

- Remove strictly dominated pure strategies
- Try all possible supports
- Find probabilities leading to indifferent payoffs

Example 3.25 (Common interest).

		Player 2 A	Player 2 B
Player 1	A	$(1,1)$	$(0,0)$
	B	$(0,0)$	$(1,1)$

We saw before that there were two pure Nash equilibriums. There are no obvious strictly dominated strategies.
Let us look for a mixed Nash equilibrium. $\sigma_1 = (x, 1-x)$ for some $x \in [0,1]$ since it is a probability vector on two states; $\sigma_2 = (y, 1-y)$ for some $y \in [0,1]$. From Corollary 3.24 (1), we get that

$$\underbrace{1 \times y}_{\text{1 plays A, 2 plays } \sigma_2} = \underbrace{1 \times (1-y)}_{\text{1 plays B, 2 plays } \sigma_2}$$

thus $y = 1/2$.
For the same reason $x = 1/2$. Thus, $(1/2, 1/2)$ for 1 and $(1/2, 1/2)$ for 2 is a Nash Equilibrium with payoff $1/2$ for both players.

Example 3.26. We continue here the example of Example 3.17:

		Player 2 A	Player 2 B	Player 2 C
Player 1	a	$(1,1)$	$(0,2)$	$(0,4)$
	b	$(0,2)$	$(5,0)$	$(1,6)$
	c	$(0,2)$	$(1,1)$	$(2,1)$

that we reduced in Example 3.19 to:

		Player 2 A	Player 2 C
Player 1	a	$(1,1)$	$(0,4)$
	c	$(0,2)$	$(2,1)$

Using the same reasoning and notations as in Example 3.25, we get for the actions of Player 1 that

$$\underbrace{1 \times y}_{\text{1 plays a}} = \underbrace{2 \times (1-y)}_{\text{1 plays c}}$$

thus $y = 2/3$.
And for the actions of Player 2:

$$\underbrace{1 \times x + 2 \times (1-x)}_{\text{2 plays A}} = \underbrace{4 \times x + 1 \times (1-x)}_{\text{2 plays C}}$$

thus $x = 1/4$.

This means that $(1/4, 3/4)$ for 1 and $(2/3, 1/3)$ for 2 is the mixed NE of the reduced game. Since strictly dominated strategies are not played, the mixed NE of the original game is $(1/4, 0, 3/4)$ for 1 and $(2/3, 0, 1/3)$ for 2.

3.3.5 *The price of anarchy*

Example 3.27 (Prisonner's dilemma again). We recall the game:

		Bob	
		Silent	Cooperate
Alice	Silent	$(-1,-1)$	$(-3,0)$
	Cooperate	$(0,-3)$	$(-2,-2)$

If we try to apply the same reasoning, we get for the actions of Alice that

$$\underbrace{-1 \times y - 3 \times (1-y)}_{\text{Alice stays silent}} = \underbrace{-2 \times (1-y)}_{\text{Alice cooperates}}$$

and we end up with $2y - 1 = 2y - 2$, that is impossible, meaning that there is no Nash Equilibrium with both actions at the same time for Alice by Corollary 3.24. Same thing occurs for Bob.
We are left with looking for Nash equilibrium with one action for both player (ie. pure NE). We already saw that Cooperate for both player was the only pure NE. It is also the mixed NE.
The payment for both players is $(-2, -2)$ which is less than the maximal payment possible of $(-1, -1)$, this is the price of anarchy.

The *price of anarchy* is the difference between the best possible action with cooperation and the Nash equilibrium.

3.3.6 *A proof of Nash's theorem*

This was done in course using a reduction to the use of Kakutani's fixed point theorem.

3.3.7 *Population games & Braess's paradox*

This was done in course as an example of population game with a high cost of anarchy.

3.4 Two player games

In this section, we focus on the important case when $N = 2$. Then the game writes in normal form $\Gamma = \{2; (\Sigma_1, \Sigma_2); (g_1, g_2)\}$.

3.4.1 Max-Mix strategies

Definition 3.28. Let $\omega \in \mathbb{R}$. We say that player i guarantees a payment ω if he has a mixed strategy that pays at least ω against any adversarial strategy:

$$\exists \sigma_i \in \Sigma_i : \forall \sigma_{-i} \in \Sigma_{-i}, \quad g_i(\sigma_i; \sigma_{-i}) \geq \omega$$

that is to say

$$\max_{\sigma_i \in \Sigma_i} \min_{\sigma_{-i} \in \Sigma_{-i}} g_i(\sigma_i; \sigma_{-i}) \geq \omega.$$

Proposition 3.29. *The maximal payoff that player i can guarantee is*

$$v_i = \max_{\sigma_i \in \Sigma_i} \min_{\sigma_{-i} \in \Sigma_{-i}} g_i(\sigma_i; \sigma_{-i}) = \max_{\sigma_i \in \Sigma_i} \min_{s_{-i} \in S_{-i}} g_i(\sigma_i; \sigma_{-i})$$

(By linearity of the payoff, the min can be taken over all actions instead of all strategies.)

Definition 3.30. A (mixed) strategy $\sigma_i \in \Sigma_i$ is *max-min* if $\min_{\sigma_{-i} \in \Sigma_{-i}} g_i(\sigma_i; \sigma_{-i}) = v_i$

A max-min policy is not necessarily a NE but it can be a sensible policy if player i is *risk-averse*, if the payoff of the other player in unknown, or if the other player is not rational.

Example 3.31.

		Player 2	
		A	B
Player 1	a	$(-15, 0)$	$(8, 1)$
	b	$(7, 0)$	$(7, 1)$

For player 1, b guarantees a payoff of 7. For player 2, B guarantees a payoff of 1.
(b,B) is a max-min equilibrium but not a Nash equilibrium.
Indeed, (a,B) is the only NE of the game. It is best if player 2 plays "well".

3.4.2 Zero-sum games

In zero sum two players games, $g_1 = -g_2 = g$.

Theorem 3.32 (Von Neumann's minimax theorem)**.** *Let Γ be a zero sum two players game with $g(\cdot, \sigma_2)$ concave for any $\sigma_2 \in \Sigma_2$ and $g(\sigma_1, \cdot)$ convex for any $\sigma_1 \in \Sigma_1$. A strategy $(\sigma_1^\star, \sigma_2^\star)$ is a (mixed) Nash Equilibrium if and only if it is max-min. Furthermore,*

$$\begin{aligned} v_1 = g(\sigma_1^\star, \sigma_2^\star) = g_1(\sigma_1^\star, \sigma_2^\star) &= \max_{\sigma_1 \in \Sigma_1} \min_{\sigma_2 \in \Sigma_2} g_1(\sigma_1; \sigma_2) \\ &= \min_{\sigma_2 \in \Sigma_2} \max_{\sigma_1 \in \Sigma_1} g_1(\sigma_1; \sigma_2) \\ &= -\max_{\sigma_2 \in \Sigma_2} \min_{\sigma_1 \in \Sigma_1} g_2(\sigma_1; \sigma_2) \\ &= -v_2. \end{aligned}$$

The payment of a Nash Equilibrium is thus $(v_1, -v_1)$; v_1 is then called the value *of the game.*

In the case of zero-sum games, finding a Nash Equilibrium amounts to finding a *saddle-point*, ie. a pair $(\sigma_1^\star, \sigma_2^\star) \in \Sigma_1 \times \Sigma_2$ such that for all $(\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$

$$g(\sigma_1, \sigma_2^\star) \leq g(\sigma_1^\star, \sigma_2^\star) \leq g(\sigma_1^\star, \sigma_2). \tag{Saddle-Point}$$

Finding a saddle point problem is a difficult optimization problem in general but, it enables to find Nash Equilibriums for zero sum games without having to manually consider all possibles supports which can get very difficult computationally when the dimension gets large.

In the next subsections, we see two cases where the (Saddle-Point) problem can be solved numerically by (variants of) usual optimization methods.

3.4.3 *The Linear case & Linear programming*

Without loss of generality, we take $S_1 = S_2 = \{1, .., n\}$, so that the space of mixed strategies is $\Sigma_1 = \Sigma_2 = \Delta_n$, the simplex in dimension n.

We consider a cost matrix $A \in \mathbb{R}^{n \times n}$ with non-negative entries so that $A_{i,j} = g(i, j)$ (with i an action of player 1 and j an action of player 2). Then, if player 1 plays x and 2 plays y (both mixed strategies in Δ_n), the payoff for player 1 is $x^\top A y$ and $-x^\top A y$ for player 2.

Since the cost is convex-concave, Von Neumann's theorem tells us that a Nash Equilibrium of this game can be obtained by solving the max min problem:

$$\max_{x \in \Delta} \min_{y \in \Delta} x^\top A y. \tag{3.1}$$

This problem is equivalent to

$$\max_{t, x \in \Delta} t \text{ such that } \min_{y \in \Delta} x^\top A y \geq t,$$

and for some real value t and $e = (1, 1, .., 1)$,

$$\begin{aligned}
&\exists x \in \Delta \text{ such that } \min_{x \in \Delta} \{x^\top A y\} \geq t \\
\Leftrightarrow &\exists x \in \mathbb{R}^n \text{ such that } x \geq 0, e^\top x = 1, \min_{i=1,..,n} \{[A^\top x]_i\} \geq t \\
\Leftrightarrow &\exists x \in \mathbb{R}^n \text{ such that } x \geq 0, e^\top x = 1, A^\top x \geq te.
\end{aligned}$$

Thus, the max min problem (3.1) is equivalent to

$$\max_{t, x} t \text{ such that } x \geq 0, e^\top x = 1, A^\top x \geq te \tag{3.2}$$

which is a linear program.

The optimum $(t^\star, x^\star)$ gives the value of the game $t^\star$ and the optimal strategy $x^\star$.

Remark 3.33 (Finding the optimal adversarial strategy). Using the same notations

$$\begin{aligned}
&\max_{x \in \Delta} \min_{y \in \Delta} x^\top A y \leq t \\
\Leftrightarrow &\min_{y \in \Delta} \max_{x \in \Delta} x^\top A y \leq t \\
\Leftrightarrow &\exists y \in \Delta \text{ such that } \max_{x \in \Delta} \{x^\top A y\} \leq t \\
\Leftrightarrow &\exists y \in \mathbb{R}^n \text{ such that } y \geq 0, e^\top y = 1, \max_{i=1,..,n} \{[Ay]_i\} \leq t \\
\Leftrightarrow &\exists y \in \mathbb{R}^n \text{ such that } y \geq 0, e^\top y = 1, Ay \leq te.
\end{aligned}$$

Thus, since (3.1) is equivalent to

$$\min_t t \text{ such that } \max_{x\in\Delta}\min_{y\in\Delta} x^\top Ay \le t,$$

it is also equivalent to

$$\min_{t,y} t \text{ such that } y \ge 0, e^\top y = 1, Ay \le te \tag{3.3}$$

which is a again a linear program whose optimal value $(t^\star, y^\star)$ gives the value of the game $t^\star$ and the optimal adversarial strategy $y^\star$. ◂

Example 3.34.

		Player 2	
		A	B
Player 1	a	$(-6,6)$	$(9,-9)$
	b	$(4,-4)$	$(-6,6)$

is a linear zero-sum game characterized by matrix $A = \begin{bmatrix} -6 & 9 \\ 4 & -6 \end{bmatrix}$

The solution of (3.2) for this game is $t^\star = 0$, $x^\star = (2/5, 3/5)$ (the solution of (3.3) is $t^\star = 0$, $y^\star = (3/5, 2/5)$).

3.4.4 *The Concave-Convex case & Extragradient*

When the payoffs are not linear, finding a saddle point

$$g(\sigma_1, \sigma_2^\star) \le g(\sigma_1^\star, \sigma_2^\star) \le g(\sigma_1^\star, \sigma_2) \tag{Saddle-Point}$$

is in general more difficult, but can still be achieved by first-order "gradient-like" methods. This kind of setup has attracted a lot of interest in the 2020's for the training of Generative Adversarial Networks (GANs).

We define $x = (\sigma_1, \sigma_2)$ and $\mathcal{X} = \Delta_n \times \Delta_n$. In this product space, we can define $v = (-\nabla_{\sigma_1} g, +\nabla_{\sigma_2} g)$ and try to move oppositely to its direction (ie. do a gradient ascent on $g(\cdot, \sigma_2)$ and a gradient descent on $g(\sigma_1, \cdot)$):

$$X_{k+1} = \mathrm{proj}_{\mathcal{X}}(X_k - \gamma v(X_k)). \tag{Gradient Descent Ascent}$$

Unfortunately, this direct strategy does not work in general.

Remark 3.35 (Failure of Gradient Descent/Ascent). Consider the problem

$$\max_{x\in\mathbb{R}}\min_{y\in\mathbb{R}} xy.$$

The only solution is $(0,0)$ but $v(x,y) = (-y, x)$ which necessarily increases the norm of (x,y). ◂

To overcome this problem, Korpelevich introduced in (**?**) the principle of *Extragradient*:

$$\begin{cases} X_{k+1/2} = \mathrm{proj}_{\mathcal{X}}(X_k - \gamma v(X_k)) \\ X_{k+1} = \mathrm{proj}_{\mathcal{X}}(X_k - \gamma v(X_{k+1/2})) \end{cases} \tag{ExtraGradient}$$

which intuitively consists in generating a *leading* point that will look forward the value of the field and apply it to the base point. This way, circular effects can be managed and convergence can be restored.

The textbook (Facchinei and Pang, 2003) gives the following result for extragradient.

Theorem 3.36 (Facchinei and Pang 2003, Th. 12.1.11)**.** *Let $\mathcal{X}$ be a closed convex set in $\mathbb{R}^n$ and v be a L-Lipschitz continuous monotone vector field on $\mathcal{X}$. Then, the iterates of Extragradient with $\gamma < 1/L$ converge to a point $X^\star$ such that*

$$\langle v(X^\star), X - X^\star \rangle \geq 0 \text{ for all } X \in \mathcal{X}.$$ [8]

[8] *We call this equation a* Variational Inequality *for v constrained to $\mathcal{X}$.*

In our situation, $\mathcal{X} = \Delta_n \times \Delta_n$ is indeed closed and convex. The vector field $v = (-\nabla_{\sigma_1} g, +\nabla_{\sigma_2} g)$ is monotone[9] since g is concave in its first argument and convex in its second. We have to add the assumption that it is L-smooth to get that v is L-Lipschitz. Then, the iterates of extragradient converge to a point $X^\star = (\sigma_1^\star, \sigma_2^\star)$ such that $\langle v(X^\star), X - X^\star \rangle \geq 0$ for all $X \in \mathcal{X}$ which is equivalent to

[9] a mapping is monotone if $\langle v(x) - v(y), x - y \rangle \geq 0$

$$\begin{aligned}
&\begin{cases} \langle -\nabla_{\sigma_1} g(\sigma_1^\star, \sigma_2^\star), \sigma_1 - \sigma_1^\star \rangle \geq 0 \\ \langle \nabla_{\sigma_2} g(\sigma_1^\star, \sigma_2^\star), \sigma_2 - \sigma_2^\star \rangle \geq 0 \end{cases} \quad \text{for all } (\sigma_1^\star, \sigma_2^\star) \in \Delta_n \times \Delta_n \\
\Leftrightarrow &\begin{cases} 0 \in -\nabla_{\sigma_1} g(\sigma_1^\star, \sigma_2^\star) + N_{\Delta_n}(\sigma_1^\star) \\ 0 \in \nabla_{\sigma_2} g(\sigma_1^\star, \sigma_2^\star) + N_{\Delta_n}(\sigma_1^\star) \end{cases} \\
\Leftrightarrow^{10} &\begin{cases} \sigma_1^\star \in \arg\max_{\Delta_n} g(\cdot, \sigma_2^\star) \\ \sigma_2^\star \in \arg\min_{\Delta_n} g(\sigma_1^\star, \cdot) \end{cases}
\end{aligned}$$

[10] Since g is concave-convex.

which is equivalent to (Saddle-Point).

Hence, we have the following result.

Corollary 3.37. *Let $g : \Delta_n \times \Delta_n \to \mathbb{R}$ be a concave-convex L-smooth payoff. Then, the iterates of Extragradient with $v = (-\nabla_{\sigma_1} g, +\nabla_{\sigma_2} g)$ and $\gamma < 1/L$ converge to a Nash Equilibrium of the corresponding zero-sum game.*

Example 3.38 (Linear example). When $g(x, y) = x^\top A y$, $v(x, y) = (-Ay, A^\top x)$. We can thus also solve linear games with this method.

Remark 3.39 ("Getting rid" of the simplex projections). The projection on the simplex is a QP that can actually be solved exactly by dynamic programming (see eg. (?)). Nevertheless, it can come out quite costly when the dimension is high.

A possibility to make these projections much easier to compute is to change the (implicit) Euclidean metric. For the simplex, an efficient example is the *Kullback-Liebler* divergence $D(x, y) = \sum_{i=1}^n x_i \log(x_i/y_i) - \sum_{i=1}^n (x_i - y_i)$, which serve as a metric on strictly positive vectors.[11]

[11] This is a particular case of *Bregman divergence* $D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y \rangle$ with $\Phi(x) = \sum_{i=1}^n x_i \log(x_i)$.

We this metric, for any positive vector y,

$$\text{proj}_{\Delta_n}^{KL}(y) = \text{argmin}_{u \in \Delta_n} D(u, y) = \frac{y}{\sum_{i=1}^n y_i} = \frac{y}{\|y\|_1}$$

which is much easier to compute!

By changing the metric of the Extragradient algorithm,[12] we obtain the *Mirror-Prox*

[12] ie. by going from $X_{k+1/2} = \text{argmin}_X \{-\gamma \langle v(X_k), X \rangle + \frac{1}{2}\|X - X_k\|^2\}$ to $X_{k+1/2} = \text{argmin}_X \{-\gamma \langle v(X_k), X \rangle + D(X, X_k)\}$

method:

$$\begin{cases} (a_{k+1/2}, b_{k+1/2}) = X_k \exp(-\gamma v(X_k)) \\ X_{k+1/2} = (\frac{a_{k+1/2}}{\|a_{k+1/2}\|_1}, \frac{b_{k+1/2}}{\|b_{k+1/2}\|_1}) \\ (a_{k+1}, b_{k+1}) = X_k \exp(-\gamma v(X_{k+1/2})) \\ X_{k+1} = (\frac{a_{k+1}}{\|a_{k+1}\|_1}, \frac{b_{k+1}}{\|b_{k+1}\|_1}) \end{cases} \quad \text{(Mirror Prox)}$$

where the exponential is performed elementwise.

The Mirror Prox method has similar theoretical guarantees but better constants, implementation, and behavior in practice. ◀

Exercises

Exercise 3.1. What is the pure Nash Equilibrium of the following game?

		Player 2	
		A	B
Player 1	a	$(3,1)$	$(2,3)$
	b	$(4,5)$	$(3,0)$
	c	$(2,2)$	$(5,4)$

Elements of Solution: (b,A) and (c,B) are pure NE. □

Exercise 3.2. What is the pure Nash Equilibrium of the following game?

		Player 2		
		A	B	C
Player 1	a	$(3,1)$	$(2,3)$	$(10,2)$
	b	$(4,5)$	$(3,0)$	$(6,4)$
	c	$(2,2)$	$(5,4)$	$(12,3)$
	d	$(5,6)$	$(4,5)$	$(9,7)$

Elements of Solution: (c,B) is the pure NE of the game. □

Exercise 3.3 (Vickrey auctions (1961))**.** Consider sealed-bid, second price auctions.[13] There are N players, and player i:

- estimates the price of the object at v_i
- its action set is $S_i = \mathbb{R}_+$ and corresponds to its bidding
- if he wins the auction (his bid is the greatest), he will make a profit based on the difference between his estimation and his bid, otherwise he will make 0 profit
- mathematically, its payoff if $g_i(s_i, s_{-i}) = v_i - \max_{j\neq i} s_j$ if $s_i > \max_{j\neq i} s_j$ and 0 else

Show that $(v_1, .., v_N)$ is a Nash Equilibrium.

[13] Such auctions are used for instance in advertisement bidding (eg. Google Ads), for mobile bandwidth acquisition (eg. FCC), etc.

Exercise 3.4 (Cournot competition)**.** Antoine Cournot (1801–1871) analyzed the spring water duopoly:

- Two firms produce an equivalent product ($N = 2$);
- Each firm decides of a production level $s_i \in \mathbb{R}$ for a cost $c_i(s_i)$;
- The selling price result from the demand vs offer, it is common to both firms and depend on the total production $p(s_1 + s_2)$.

The profit/payoff for company 1 is $g_1(s_1, s_2) = s_1 p(s_1 + s_2) - c_1(s_1)$; the one for company 2 is $g_2(s_1, s_2) = s_2 p(s_1 + s_2) - c_2(s_2)$.
Which quantity to produce? Compute the explicit production when $c_i(s_i) = c_i \times s_i$ for some $c_i > 0$ and $p(u) = \max(a - b \times u; 0)$ for $a, b > 0$.

Exercise 3.5. Three companies are in concurrence in a product. They choose their price p_i , $i = 1, .., 3$, simultaneously (a positive real number). The customer demand is then given by $q_i = 100 - 3p_i + \sum_{j \neq i} p_j$ for each company $i = 1, .., 3$. The reward of company i is then $p_i \times q_i$.

1. What are the actions and cost functions of each player?
2. What is the best response of player 1 to a strategy $(p_2, p_3) \in \mathbb{R}_+ \times \mathbb{R}_+$ of the two other players?
3. What is the Nash Equilibrium of the game?

Exercise 3.6. Suppose that 4000 drivers want to go from town A to town B. To do so they have two alternatives:

i) going from town A to town C which takes 45 minutes, then from town C to town B which takes u/100 minutes where u is the number of vehicles that take this road;

ii) going from town A to town D which takes v/100 minutes where v is the number of vehicles that take this road, then from town D to town B which takes 45 minutes.

To improve the road network, public authorities plan to add a (bidirectional) road between C and D that takes 0 minutes. Compute Nash Equilibria in both cases and conclude. What is the price of anarchy here?

Exercise 3.7. Depending on the value of the parameter $x \in \mathbb{R}$, give the pure and mixed Nash Equilibria for the following game:

		Player 2	
		A	B
Player 1	A	$(0.5, 0.5)$	$(x, 1-x)$
	B	$(1-x, x)$	$(0.5, 0.5)$

Exercise 3.8. We consider the following game in normal form:

		Player 2	
		G	D
Player 1	G	$(0, 2)$	$(3, 0)$
	D	$(2, 1)$	$(1, 3)$

1. What are the actions and cost functions of each player?
2. Is there a Nash Equilibrium with only pure strategies?
3. What are all Nash Equilibria with mixed strategies?

Exercise 3.9. Consider the following two-player game in normal form. Both players

have the strategy set $\{A, B\}$ and the payoff matrix is given by

	A	B
A	$(3,3)$	$(0,2)$
B	$(2,0)$	$(1,1)$

(a) Identify all *pure strategy* Nash equilibria.

(b) Find the *mixed strategy* Nash equilibrium.

Elements of Solution:

(a) **Pure Strategies:**

- If both players play A, neither can improve by deviating (since deviating to B would give a payoff of 2 instead of 3). Hence, (A, A) is a Nash equilibrium.
- If both play B, unilateral deviation is not profitable (switching from B to A would lower a player's payoff from 1 to 0). Thus, (B, B) is also a Nash equilibrium.

(b) **Mixed Strategy Equilibrium:** Denote by p (resp. q) the probability that Player 1 (resp. Player 2) plays A. For a mixed equilibrium, each player must be indifferent between playing A and B.

Player 1:

$$\begin{aligned}\text{Payoff from } A &= 3q + 0\,(1-q) = 3q,\\ \text{Payoff from } B &= 2q + 1\,(1-q) = 2q + 1 - q = 1 + q.\end{aligned}$$

Set equal for indifference:

$$3q = 1 + q \quad \Longrightarrow \quad 2q = 1 \quad \Longrightarrow \quad q = \frac{1}{2}.$$

Player 2: Similarly, denote by p the probability that Player 1 plays A. Then

$$\begin{aligned}\text{Payoff from } A &= 3p + 0\,(1-p) = 3p,\\ \text{Payoff from } B &= 2p + 1\,(1-p) = 2p + 1 - p = 1 + p.\end{aligned}$$

For indifference:

$$3p = 1 + p \quad \Longrightarrow \quad 2p = 1 \quad \Longrightarrow \quad p = \frac{1}{2}.$$

Hence, the unique mixed Nash equilibrium is

$$\left(p = \frac{1}{2},\ q = \frac{1}{2}\right),$$

in which each player randomizes equally between A and B.

□

Exercise 3.10. Consider the following game between two players. The strategy sets are:

$$\text{Player 1} : \{A, B\} \quad \text{and} \quad \text{Player 2} : \{X, Y\},$$

with payoff matrix

	X	Y
A	$(4, 3)$	$(0, 1)$
B	$(2, 0)$	$(2, 4)$

(a) Determine the pure strategy Nash equilibria.

(b) Find the mixed strategy Nash equilibrium and show that at equilibrium each player is *strongly indifferent* among the pure strategies in the support of their mixed strategy.

Elements of Solution:

(a) **Pure Strategies:**

- For Player 1: If Player 2 plays X, A gives 4 and B gives 2; if Player 2 plays Y, A gives 0 and B gives 2.
- For Player 2: If Player 1 plays A, X gives 3 and Y gives 1; if Player 1 plays B, X gives 0 and Y gives 4.

Thus, the best responses are:

- (A, X): If Player 1 plays A, Player 2's best response is X, and if Player 2 plays X, Player 1's best response is A.
- (B, Y): If Player 1 plays B, Player 2's best response is Y, and if Player 2 plays Y, Player 1's best response is B.

Hence, the pure Nash equilibria are (A, X) and (B, Y).

(b) **Mixed Strategy Equilibrium:** Let Player 1 play A with probability p (and B with $1 - p$), and Player 2 play X with probability q (and Y with $1 - q$). For each player to mix, they must be indifferent between their strategies.

For Player 1:

$$\text{Payoff from } A = 4q + 0\,(1 - q) = 4q,$$

$$\text{Payoff from } B = 2q + 2\,(1 - q) = 2q + 2 - 2q = 2.$$

Indifference implies

$$4q = 2 \quad \Longrightarrow \quad q = \frac{1}{2}.$$

For Player 2:

$$\text{Payoff from } X = 3p + 0\,(1 - p) = 3p,$$

$$\text{Payoff from } Y = 1p + 4\,(1 - p) = p + 4 - 4p = 4 - 3p.$$

Equate these for indifference:

$$3p = 4 - 3p \quad \Longrightarrow \quad 6p = 4 \quad \Longrightarrow \quad p = \frac{2}{3}.$$

Thus, the unique mixed equilibrium is:

$$\left(p = \frac{2}{3},\ q = \frac{1}{2}\right).$$

Strong Indifference: In this equilibrium the expected payoffs are:

$$\text{For Player 1:} \qquad 4q = 4\left(\frac{1}{2}\right) = 2, \quad \text{and} \quad 2 = 2.$$

$$\text{For Player 2:} \qquad 3p = 3\left(\frac{2}{3}\right) = 2, \quad \text{and} \quad 4 - 3p = 4 - 2 = 2.$$

Since each player's pure strategies (that are played with positive probability) yield the same expected payoff, they are strongly indifferent among them.

□

Exercise 3.11. Consider the following bimatrix game between Player 1 (rows) and Player 2 (columns). Their available strategies are

$$\text{Player 1: } R_1,\ R_2,\ R_3, \quad \text{Player 2: } C_1,\ C_2,\ C_3.$$

The payoff matrix (written as (u_1, u_2)) is:

	C_1	C_2	C_3
R_1	(3, 2)	(2, 1)	(0, 1)
R_2	(2, 1)	(3, 3)	(1, 0)
R_3	(1, 0)	(1, 1)	(0, −1)

(a) Show that for Player 1 the strategy R_3 is strictly dominated by R_2, and for Player 2 the strategy C_3 is strictly dominated by C_1. (Hint: Compare payoffs column by column.)

(b) After eliminating R_3 and C_3, find all pure-strategy Nash equilibria of the resulting reduced 2×2 game.

(c) In the reduced game, compute the mixed-strategy Nash equilibrium and verify that it makes each player *strongly indifferent* between the strategies in the support.

Elements of Solution:

(a) **For Player 1:**

- When Player 2 plays C_1: R_1 gives 3, R_2 gives 2, and R_3 gives 1.
- When Player 2 plays C_2: R_1 gives 2, R_2 gives 3, and R_3 gives 1.
- When Player 2 plays C_3: R_1 gives 0, R_2 gives 1, and R_3 gives 0.

In every column R_2 yields a higher payoff than R_3 (and in the C_3 column, $1 > 0$). Hence, R_3 is strictly dominated by R_2.

For Player 2:

- When Player 1 plays R_1: C_1 gives 2, C_2 gives 1, C_3 gives 1.
- When Player 1 plays R_2: C_1 gives 1, C_2 gives 3, C_3 gives 0.
- (After elimination, Player 1 will never play R_3.)

Comparing C_1 and C_3: against R_1, $2 > 1$; against R_2, $1 > 0$. Hence C_3 is strictly dominated by C_1.

(b) After eliminating R_3 and C_3, the reduced game is:

	C_1	C_2
R_1	$(3, 2)$	$(2, 1)$
R_2	$(2, 1)$	$(3, 3)$

Best responses:

- For Player 1: If Player 2 plays C_1, best response is R_1 (3 vs. 2); if Player 2 plays C_2, best response is R_2 (3 vs. 2).
- For Player 2: If Player 1 plays R_1, best response is C_1 (2 vs. 1); if Player 1 plays R_2, best response is C_2 (3 vs. 1).

Hence, the pure Nash equilibria of the reduced game are (R_1, C_1) and (R_2, C_2).

(c) **Mixed-Strategy Equilibrium:** Denote by p the probability that Player 1 plays R_1 (and $1 - p$ for R_2), and by q the probability that Player 2 plays C_1 (and $1 - q$ for C_2).

Player 1's indifference:

$$U_1(R_1) = 3q + 2(1 - q) = 2 + q,$$
$$U_1(R_2) = 2q + 3(1 - q) = 3 - q.$$

Setting $2 + q = 3 - q$ yields $2q = 1$, so $q = \frac{1}{2}$.

Player 2's indifference:

$$U_2(C_1) = 2p + 1(1 - p) = 1 + p,$$
$$U_2(C_2) = 1p + 3(1 - p) = 3 - 2p.$$

Setting $1 + p = 3 - 2p$ gives $3p = 2$, so $p = \frac{2}{3}$.

Thus, the unique mixed-strategy Nash equilibrium in the reduced game is

$$\left(p = \frac{2}{3},\ q = \frac{1}{2}\right).$$

Verify the indifference:

$$U_1(R_1) = 2 + \frac{1}{2} = 2.5, \quad U_1(R_2) = 3 - \frac{1}{2} = 2.5,$$

and

$$U_2(C_1) = 1 + \frac{2}{3} \approx 1.67, \quad U_2(C_2) = 3 - 2 \cdot \frac{2}{3} \approx 1.67.$$

In equilibrium each player obtains the same expected payoff from any strategy played with positive probability; hence, they are strongly indifferent among the strategies in their support.

□

Exercise 3.12 (Cournot Duopoly with Continuous Quantities)**.** Consider two firms (Firm 1 and Firm 2) competing in a Cournot duopoly. The market inverse demand function is

$$P(Q) = a - Q, \quad \text{with } Q = q_1 + q_2,$$

and both firms have constant marginal cost c (with $0 < c < a$). The profit functions are

$$\pi_i(q_1, q_2) = q_i\Big(a - q_1 - q_2 - c\Big), \quad i = 1, 2.$$

(a) Derive the best response function for each firm.

(b) Find the Nash equilibrium quantities (q_1^*, q_2^*).

(c) Determine the equilibrium market price.

Elements of Solution:

(a) **Best Response Functions:**

For Firm 1, fix q_2 and maximize

$$\pi_1(q_1, q_2) = q_1(a - q_1 - q_2 - c).$$

Differentiating with respect to q_1 gives:

$$\frac{\partial \pi_1}{\partial q_1} = a - q_1 - q_2 - c - q_1 = a - c - q_2 - 2q_1.$$

Setting this derivative equal to zero:

$$a - c - q_2 - 2q_1 = 0 \quad \Longrightarrow \quad q_1 = \frac{a - c - q_2}{2}.$$

Similarly, by symmetry for Firm 2:

$$q_2 = \frac{a - c - q_1}{2}.$$

(b) **Nash Equilibrium Quantities:**

Substitute Firm 2's best response into Firm 1's:

$$q_1 = \frac{a - c - \frac{a - c - q_1}{2}}{2}.$$

Multiply numerator and denominator appropriately:

$$q_1 = \frac{2(a - c) - (a - c - q_1)}{4} = \frac{(a - c) + q_1}{4}.$$

Multiply both sides by 4:

$$4q_1 = a - c + q_1 \quad \Longrightarrow \quad 3q_1 = a - c.$$

Hence,

$$q_1^* = \frac{a - c}{3}.$$

By symmetry,

$$q_2^* = \frac{a - c}{3}.$$

(c) **Equilibrium Price:**

The total equilibrium quantity is:

$$Q^* = q_1^* + q_2^* = \frac{a-c}{3} + \frac{a-c}{3} = \frac{2(a-c)}{3}.$$

Thus the equilibrium price is:

$$P^* = a - Q^* = a - \frac{2(a-c)}{3} = \frac{3a - 2a + 2c}{3} = \frac{a+2c}{3}.$$

□

Exercise 3.13 (Subgame-Perfect Equilibrium in a Sequential Game)**.** Consider the following extensive-form game between Player 1 and Player 2:

- First, Player 1 chooses between actions A and B.
- If Player 1 chooses A, then Player 2 chooses between C and D.
- If Player 1 chooses B, then Player 2 chooses between E and F.

The payoffs (written as (u_1, u_2)) are given by:

If A is chosen:	C	D
	$(3,2)$	$(1,4)$

If B is chosen:	E	F
	$(5,0)$	$(0,0)$

(a) Using backward induction, determine the optimal action for Player 2 in each subgame.

Definition 3.40 (Subgame-Perfect Equilibrium)**.** A strategy profile in an extensive-form game is a *subgame-perfect equilibrium* (SPE) if it induces a Nash equilibrium in every subgame of the original game. In other words, the strategy profile is obtained by applying backward induction so that at every decision node the players' actions are optimal given the continuation of the game.

(b) Find the subgame-perfect Nash equilibrium (SPE) of the game and state the outcome.

Elements of Solution:

(a) **Subgames:**

- *Subgame after A:* Player 2 chooses between:

$$C: \ u_2 = 2, \quad D: \ u_2 = 4.$$

Hence, Player 2's optimal action is D.

- *Subgame after B:* Player 2 chooses between:

$$E: \ u_2 = 0, \quad F: \ u_2 = 0.$$

(Player 2 is indifferent; assume she chooses E by convention.)

(b) **Backward Induction:** Given Player 2's responses:

- If Player 1 chooses A, the outcome is (A, D) with payoff $(1, 4)$.
- If Player 1 chooses B, the outcome is (B, E) with payoff $(5, 0)$.

Since Player 1 prefers a payoff of 5 over 1, her optimal action is to choose B. Thus, the SPE is:

Player 1: Choose B,
Player 2: If A is reached, choose D; if B is reached, choose E.

The equilibrium outcome is (B, E) with payoffs $(5, 0)$.

□

CHAPTER 4 Optimization & Robustness in Measure spaces

Now that we have seen a few examples of robustness in the Euclidean setting, we now turn to robustness in measure spaces. We will pay a special attention to stochastic optimization problems, and relate them to linear optimization in measure spaces. Finally, we will survey interesting notions of distance and divergence for probability measures.

4.1 Stochastic optimization & Robustness

Let us consider the problem of minimizing in x a smooth function that also depends on a random uncertainty:

$$\min_x f(x; X)$$

then, the problem is ill-posed. Indeed, the optimality conditions in x are $\nabla_x f(x; X) = 0$ ... for almost all X? in expectation?

Instead of making the problem more precise right now, let us mimicking what would be done without the random component. This way, we see what happens for an *unseen randomness* (something that happens in practice...).

A first thing to notice is that if μ has variance 0, we are in the standard case of (deterministic) optimization without any robustness problem. So variance plays an important role...

4.1.1 From Stochastic Gradient Descent to Stochastic Objectives

Let write a gradient algorithm see what happens! Let us start at some x_0 and for all k, iterate

$$x_{k+1} = x_k - \gamma_k \nabla_x f(x_k; X_{k+1}) \tag{4.1}$$

where it is important that we observe X_{k+1}, that is $\mathcal{F}_{k+1}$-measurable[14] but not $\mathcal{F}_k$-measurable, when computing the gradient at x_k. This algorithm is classically called *Stochastic Gradient Descent (SGD)*.

[14] $\mathcal{F}_k$ denotes the natural filtration, i.e., the sigma algebra generated by x_0, ..., x_k.

Taking the expectation, we get

$$\mathbb{E}\left[x_{k+1}|\mathcal{F}_k\right] = x_k - \gamma_k \mathbb{E}\left[\nabla_x f(x_k; X_{k+1})|\mathcal{F}_k\right]$$

and we would love to exchange expectation and integral...

Lemma 4.1. *Let $\mathcal{X}$ be a measure space and suppose that the function $f : \mathbb{R}^n \times \mathcal{X} \to \mathbb{R}$ satisfies the following conditions:*

(*a*) Differentiability: *$f(\cdot; X)$ is C^1 for all $X \in \mathcal{X}$.*

(*b*) Smoothness: *$\nabla_x f(\cdot; X)$ is L-Lipschitz for all $X \in \mathcal{X}$.*

(*c*) Integrability: *$f(x; \cdot)$ and $\nabla_x f(x; \cdot)$ are integrable with respect to μ for a certain fixed $x \in \mathbb{R}^n$*

Then, the function $F : x \mapsto \mathbb{E}[\nabla_x f(x; X)]$ is differentiable and for all x, $\mathbb{E}[\nabla_x f(x; X)] = \nabla_x F(x)$.

Proof. We may assume without loss of generality that both $f(0; X)$ and $\nabla_x f(0; X)$ are integrable thanks to condition (*c*). Consider the function

$$g\colon (x; X) \mapsto \frac{f(x; X)}{\|x\|^2 + 1}.$$

Since the gradient of f is L-Lipschitz in x by condition (*b*), we have using the descent lemma (see Lemma A.14) that

$$|f(x; X) - f(0; X)| \leq \|\nabla_x f(0; X)\|\|x\| + \frac{L}{2}\|x\|^2$$

so that g is upper bounded by an integrable function uniformly in x as

$$|g(x; X)| \leq |f(0; X)| + \|\nabla_x f(0; X)\| + \frac{L}{2}. \tag{4.2}$$

We also have

$$\begin{aligned}\nabla_x g(x; X) &= \nabla_x f(x; X)\frac{1}{\|x\|^2+1} - x\frac{2f(x; X)}{(\|x\|^2+1)^2} = \nabla_x f(x; X)\frac{1}{\|x\|^2+1} - x\frac{2g(x; X)}{\|x\|^2+1} \\ &= \nabla_x f(0; X)\frac{1}{\|x\|^2+1} + (\nabla_x f(x; X) - \nabla_x f(0; X))\frac{1}{\|x\|^2+1} - x\frac{2g(x; X)}{\|x\|^2+1}\end{aligned}$$

Using again Lipschitz continuity of the gradient of f, $\nabla_x g(x; X)$ is upper bounded by an integrable function, uniformly in x, as

$$\begin{aligned}\|\nabla_x g(x; X)\| &\leq \|\nabla_x f(0; X)\| + L + 2g(x; X) \\ &\leq 3\|\nabla_x f(0; X)\| + 2L + 2|f(0; X)|.\end{aligned} \tag{4.3}$$

Hence, we have that i) $\nabla_x g(x; X)$ exists for all x (as f is C^1) and ii) both $X \mapsto g(x; X)$ and $X \mapsto \nabla_x g(x; X)$ are bounded by functions in $L^1(\mu)$ uniformly in x thanks to Eqs. (4.2) and (4.3) since $|f(0; X)|$ and $\|\nabla_x f(0; X)\|$ belong to $L^1(\mu)$. Hence, we have the appropriate domination assumptions to differentiate under the integral for the function g so that for all x, the function $G : x \mapsto \mathbb{E}[g(x; X)]$ is differentiable and $\nabla_x G(x) = \mathbb{E}[\nabla_x g(x; X)]$ (see e.g. (Folland, 1999, Th. 2.27)).
Now, turning back to f, since for all x, $f(x; X) = g(x; X)(\|x\|^2 + 1)$, let $F(x) = \mathbb{E}[\nabla_x f(x; X)] = G(x)(\|x\|^2 + 1)$ and thus $\nabla_x F(x) = \nabla_x G(x)(\|x\|^2 + 1) + 2xG(x)$. Also, for all x

$$\nabla_x f(x; X) = \nabla_x g(x; X)(\|x\|^2 + 1) + 2xg(x; X)$$

whose right hand side is integrable as shown above. This enables us to conclude that for all x,

$$\begin{aligned}\mathbb{E}[\nabla_x f(x; X)] &= \mathbb{E}[\nabla_x g(x; X)](\|x\|^2 + 1) + 2x\mathbb{E}[g(x; X)] \\ &= \nabla_x G(x)(\|x\|^2 + 1) + 2xG(x) = \nabla_x F(x)\end{aligned}$$

which is the claimed result. □

Using this result, we get that provided that the (X_k) are iid., we have

$$\mathbb{E}\left[x_{k+1}|\mathcal{F}_k\right] = x_k - \gamma_k \nabla_x F(x_k)$$

and thus, in expectation, a step of gradient on our stochastic objective is a gradient step on the average objective. It is thus legitimate to investigate problems of the form

$$\min_x F(x) := \mathbb{E}_{X\sim\mu}\left[f(x;X)\right] \tag{SP}$$

in view of algorithms of the form (4.1). For this, the convex and non-convex cases are very different in terms of results.

Remark 4.2 (Convergence of the stochastic gradient descent algorithm (4.1)). Let X be a measure space and suppose that the functions $f : \mathbb{R}^n \times \mathrm{X} \to \mathbb{R}$ and $F : x \mapsto \mathbb{E}[\nabla_x f(x;X)]$ satisfy the following conditions:

(*a*) *Differentiability:* $f(\cdot;X)$ is C^1 for all $X \in \mathrm{X}$.

(*b*) *Smoothness:* $\nabla_x F$ is L-Lipschitz

(*c*) *Noise:* The sequence (X_k) are iid. and $\mathbb{E}\left[\|\nabla_x f(x;X) - \mathbb{E}[\nabla_x f(x;X)]\|^2\right] \leq \sigma^2$ for all x with $\sigma < +\infty$.

The convergence results for the iterates (4.1) are quite different in the convex and general case. In the convex case, if $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$, then x_k converges almost surely to a minimizer of F. This is the topic of Exercise 4.2. Apart from these results, we also have ones on the asympotic normality (Fabian, 1968) or extensions to infinite variance (Wang et al., 2021).

In the non-convex case, the strategy is in general to show that asympotically the iterates get close to the (continuous-time) gradient flow, as detailed in (Benaïm, 2006, Chap. 3,4). ◀

We thus see that classical methods of stochastic approximation natively lead to a noisy minimization of the *expected* function. This means that

- the uncertainty is handled in average
- the samples (X_k) have to be draw exactly from the distribution μ

Example 4.3 (Convergence and Γ-convergence). We could say that if two distributions are close, the expected functions are close and then the values and minimizers are close.
The first part is true, indeed the weak convergence of a sequence of positive probability measures (μ_n) to a probability measure μ is equivalent to having $\mathbb{E}_{X\sim\mu_n}\left[f(x;X)\right] \to \mathbb{E}_{X\sim\mu}\left[f(x;X)\right] = F(x)$. We can thus hope to have a pointwise convergence of the objective.
Nevertheless, even though the values will be close, this mode of convergence does not imply a convergence of the minimizers. For instance, let $F_n(x) = (x-n)^2/n^n$, then (F_n) converges pointwise to 0 but the minimizers diverge. And this is in the convex case ! For the non convex case, we can design camel humps that have the same minimizer converging to one that has the other minimizer.
This is a important difference between converge and Γ-convergence of functions.[15]

[15] Γ-convergence of a sequence of functions is equivalent to the Kuratowski convergence of their epigraphs. Formally, (F_n) Γ-converges to F if for every sequence x_n such that $x_n \to x$, $F(x) \leq \liminf_{n\to\infty} F_n(x_n)$ and for every x, there is a sequence x_n converging to x $F(x) \geq \limsup_{n\to\infty} F_n(x_n)$.

4.1.2 Robustness in Stochastic Optimization

In the same flavor as robust optimization, distributionally robust optimization proposes to solve

$$\min_{x} \sup_{\nu \in \mathcal{U}(\mu)} \mathbb{E}_{X\sim\nu}\left[f(x;X)\right] \tag{DRO}$$

where $\mathcal{U}(\mu)$ is a neighborhood of the distribution μ in the space of distributions.

This approach has been investigated for a long time but is revitalized presently for its applications in machine learning. A chapter will be devoted to it but first we need som prerequisites on how to compare distributions, which will be covered in Section 4.3.

Remark 4.4 (Application to Statistics). In a statistical perspective, let $\mu = \mu_n$ be an empirical distribution i.e., $\mu_n = \frac{1}{n}\sum_{i=1}^{N} Y_i$ where the (Y_i) are draw independently from distribution μ^0.

In terms of notations, k will be an iteration counter, related to a sample X_k from $\mu = \mu_n$ while n will represent the number of samples in the empirical distribution μ_n.

Obviously,

$$\mathbb{E}_{X\sim\mu}\left[f(x;X)\right] = \frac{1}{n}\sum_{i=1}^{n} f(x;Y_i)$$

is equal to $\mathbb{E}_{X\sim\mu^0}\left[f(x;X)\right]$ *in average* but again, the variance of the objective is essential to control the robustness of objectives.

If $n \to \infty$, the *uniform law of large numbers* states that $\mathbb{E}_{X\sim\mu}\left[f(x;X)\right]$ converges in probability to $\mathbb{E}_{X\sim\mu^0}\left[f(x;X)\right]$ pointwise (uniformly in some compact) in x,[16] this is at the heart of the theory of M-estimators.

[16] provided that f is continuous in x and uniformly intergrable, with a dominating functions, ...

If n is fixed, then finite-sample concentration bounds on the distribution can be used but we are back to the previous problem of stability. A solution to avoid this is to consider a robust problem on the distributions. This will be the topic of Chapter 5. ◀

4.1.3 Stability in Stochastic Optimization

In the (DRO) objective above, one can notice that under mild assumptions, the objective is linear. We can thus ask ourselves the same questions as in Section 2.1: under which conditions are the solutions of the problem

$$\sup_{\nu\in\mathcal{U}} \mathbb{E}_{X\sim\nu}\left[f(x;X)\right]$$

on the border of $\mathcal{U}$? At extremal points of $\mathcal{U}$? And what does it mean to be an extremal point of $\mathcal{U}$?

In general, it is difficult to say but when $\mathcal{U}$ is linear, we have additional information. Indeed, we know that finite-dimensional linear program on $\mathbb{R}^d_+$ with m equality constraints admits solutions with at most m non-null coordinates (See **??**).

Here, we can consider the following analogue. Let Z be a closed convex subset of a Euclidean space, and let $\mathcal{P}(Z)$ denote the set of probability measures on Z. Let $\psi, \phi_1, \ldots, \phi_m : Z \to \mathbb{R}$ be measurable functions and $v \in \mathbb{R}^m$. Consider now the

optimization problem

$$\begin{array}{ll} \min_\gamma & \displaystyle\int_Z \psi(z)\,\mathrm{d}\gamma(z) \\ \text{subject to} & \gamma \in \mathcal{P}(Z) \\ & \displaystyle\int_Z \phi_i(z)\,\mathrm{d}\gamma(z) = v_i \quad \forall i = 1,\dots,m, \end{array} \tag{4.4}$$

and denote by $\mathcal{F}$ the feasible region of (4.4) and by $\mathrm{ext}(\mathcal{F})$ the set of extreme points of $\mathcal{F}$.

Theorem 4.5. *Suppose that for all $\gamma \in \mathcal{F}$, at least one of the integrals $\int_X [\psi(z)]_+ \,\mathrm{d}\gamma(z)$ and $\int_X [-\psi(z)]_+ \,\mathrm{d}\gamma(z)$ is finite and that $\int_Z |\phi_i|(z)\,\mathrm{d}\gamma(z) < \infty$ for all $i = 1,\dots,m$. Then,*

$$\begin{aligned} \sup\left\{\int_Z \psi(z)\,\mathrm{d}\gamma(z) : \gamma \in \mathcal{F}\right\} &= \sup\left\{\int_Z \psi(z)\,\mathrm{d}\gamma(z) : \gamma \in \mathrm{ext}(\mathcal{F})\right\} \\ &= \sup\left\{\int_Z \psi(z)\,\mathrm{d}\gamma : \gamma \in \mathcal{F} \cap \mathcal{D}_{m+1}(Z)\right\}, \end{aligned}$$

where $\mathcal{D}_{m+1}(Z)$ is the set of non-negative discrete measures supported on at most $m+1$ points in Z.

Before proving this result, we need some machinery on Langrangian duality in general Banach spaces.

4.2 Convex Optimization with Linear Constraints in Banach spaces

In this section, we extend the classical Lagrangian machinery to infinite dimensional Banach spaces. We only go through the results and refer to (Peypouquet, 2015, Th. 3.66), and more generally (Bonnans and Shapiro, 2013; Peypouquet, 2015; Rockafellar, 1970) for textbooks on the topic.

4.2.1 Lagrangian duality

Here, one can take m constraints, and thus $Y = \mathbb{R}^m$

Let X be a Banach space. Assume:

- $f : X \to \mathbb{R}$ is convex and Fréchet differentiable,
- $A : X \to Y$ is a bounded linear operator into a Banach space Y that is partially ordered by a closed convex cone $K \subset Y$ (so that the inequality $Ax \leq b$ is defined as $Ax \in b - K$),
- $C : X \to Z$ is a bounded linear operator into a Banach space Z,
- $b \in Y$ and $d \in Z$.

The infinite-dimensional convex optimization problem with linear constraints is

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{subject to} & Ax \leq b, \\ & Cx = d. \end{array}$$

The Lagrangian is defined by

$$L(x, \lambda, \mu) = f(x) + \langle \lambda, Ax - b \rangle_Y + \langle \mu, Cx - d \rangle_Z,$$

where

- $\lambda \in Y^*$ (the dual of Y) and the dual pairing $\langle \lambda, y \rangle_Y$ is used; moreover, the dual variable satisfies $\lambda \geq 0$ (i.e. $\langle \lambda, k \rangle_Y \geq 0$ for all $k \in K$),
- $\mu \in Z^*$.

The dual function is

$$g(\lambda, \mu) = \inf_{x \in X} L(x, \lambda, \mu).$$

Under convexity and suitable regularity conditions, including an appropriate infinite-dimensional version of Slater's condition (i.e., there exists $\bar{x} \in X$ such that $A\bar{x} < b$ in the sense that $b - A\bar{x}$ belongs to the interior of K and $C\bar{x} = d$), strong duality holds.

Theorem 4.6 (KKT Conditions for Infinite-Dimensional Convex Problems with Linear Constraints)**.** *Assume that*

1. $f : X \to \mathbb{R}$ *is convex and Fréchet differentiable,*
2. $A : X \to Y$ *and* $C : X \to Z$ *are bounded linear operators,*
3. *Slater's condition holds: there exists* $\bar{x} \in X$ *such that* $A\bar{x} \in b - \operatorname{int}(K)$ *and* $C\bar{x} = d$.

Then $x^* \in X$ *is optimal if and only if there exist multipliers* $\lambda^* \in Y^*$ *(with* $\lambda^* \geq 0$*) and* $\mu^* \in Z^*$ *such that:*

1. ***Stationarity:***
$$Df(x^*) + A^*\lambda^* + C^*\mu^* = 0 \quad \text{in } X^*,$$
where A^* *and* C^* *are the adjoints of* A *and* C*, respectively.*
2. ***Primal Feasibility:***
$$Ax^* \in b - K, \quad Cx^* = d.$$
3. ***Dual Feasibility:***
$$\lambda^* \geq 0.$$
4. ***Complementary Slackness:***
$$\langle \lambda^*, Ax^* - b \rangle_Y = 0.$$

Sketch of proof. **Necessity:** Assume that x^* is an optimal solution. Under the infinite-dimensional version of Slater's condition, strong duality holds (see, e.g., (Rockafellar, 1970) and (Bonnans and Shapiro, 2013)). Thus, there exists a saddle point (x^*, λ^*, μ^*) of the Lagrangian. In particular, for all $x \in X$ and for all (λ, μ) with $\lambda \geq 0$,

$$L(x^*, \lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*) \quad \text{and} \quad L(x^*, \lambda^*, \mu^*) \geq L(x^*, \lambda, \mu).$$

Since x^* minimizes $L(\cdot, \lambda^*, \mu^*)$ and f is Fréchet differentiable, the first-order necessary condition gives

$$Df(x^*) + A^*\lambda^* + C^*\mu^* = 0.$$

Primal feasibility follows from the problem statement. Dual feasibility $\lambda^* \geq 0$ is built into the dual problem. Complementary slackness follows from the saddle point property. In fact, if there were any nonzero dual pairing $\langle \lambda^*, Ax^* - b \rangle_Y$ (with $Ax^* - b \in -K$), then one could improve the dual objective, contradicting the saddle point property.

Sufficiency: Conversely, assume that there exist x^*, $\lambda^* \geq 0$, and μ^* satisfying the above KKT conditions. For any feasible x (i.e., $Ax \in b - K$ and $Cx = d$), convexity of f and the saddle point property of the Lagrangian imply that

$$f(x) \geq L(x, \lambda^*, \mu^*) \geq L(x^*, \lambda^*, \mu^*) = f(x^*),$$

where the last equality uses stationarity, primal feasibility, and complementary slackness. Thus, x^* is optimal.
A rigorous justification of these steps in the infinite-dimensional setting is given in (Rockafellar, 1970, Chapter 5) and (Bonnans and Shapiro, 2013). □

Remark 4.7 (Optimization on probabilities versus measures). tl:dr : optimize over (positive) measures for topological reasons (to be able to define the dual), add the unit mass demand of probabilities to the constraints. ◀

4.2.2 Fenchel duality

Let E be a real locally convex topological vector space with continuous dual E^*. Let $\Phi : E \to (-\infty, +\infty]$ be an extended real-valued functional.

Definition 4.8 (Fenchel–Legendre conjugate)**.** The Fenchel conjugate of Φ is the functional

$$\Phi^* : E^* \to (-\infty, +\infty] \quad \text{defined by} \quad \Phi^*(\ell) = \sup_{f \in E} \{\langle \ell, f \rangle - \Phi(f)\},$$

where $\langle \ell, f \rangle$ denotes the dual pairing between E^* and E.

Lemma 4.9. *For any functional $\Phi : E \to (-\infty, +\infty]$:*

1. *Φ^* is convex and lower semicontinuous on E^*.*
2. *Φ^* is proper if and only if Φ is not identically $+\infty$.*
3. *The mapping $\Phi \mapsto \Phi^*$ is order-reversing.*

Proof. Convexity follows from taking a supremum of affine functions. Lower semicontinuity follows from the fact that Φ^* is a supremum of continuous linear functionals. The remaining statements follow directly from the definition. □

Theorem 4.10 (Fenchel–Moreau)**.** *Let E be a locally convex topological vector space and $\Phi : E \to (-\infty, +\infty]$ be proper, convex, and lower semicontinuous. Then*

$$\Phi^{**} = \Phi,$$

where

$$\Phi^{**}(f) = \sup_{\ell \in E^*} \{\langle \ell, f \rangle - \Phi^*(\ell)\}.$$

Proof. This is the classical Fenchel–Moreau theorem. The inequality $\Phi^{**} \leq \Phi$ holds for all functionals. If Φ is convex and lower semicontinuous, then the epigraph of Φ is closed and convex, and separation arguments yield $\Phi^{**} = \Phi$. □

Duality in Function Spaces

Let $X \subset \mathbb{R}^n$ be a closed (hence locally compact) subset.

- If $E = C_b(X)$, then $E^* = \mathcal{M}(X)$ (Radon measures).
- If $E = L^p(X)$ with $1 < p < \infty$, then $E^* = L^q(X)$.

- If $E = C_0(X)$, then $E^* = \mathcal{M}(X)$.

In each case, the Fenchel conjugate takes the form

$$\Phi^*(\mu) = \sup_{f \in E} \left\{ \int f \, d\mu - \Phi(f) \right\}.$$

4.2.3 *Duality in measure and function spaces*

4.2.4 *Proof of Theorem 4.5*

We will rely on a result on *measures* as motivated by Remark 4.7. Let Z be a closed convex subset of a Euclidean space, and let $\mathcal{M}_+(Z)$ denote the set of positive measures on Z. Let $\psi, \phi_1, \ldots, \phi_m : Z \to \mathbb{R}$ be measurable functions and $v \in \mathbb{R}^m$. Consider now the optimization problem

$$\begin{array}{ll} \min_\gamma & \displaystyle\int_Z \psi(z)\, \mathrm{d}\gamma(z) \\ \text{subject to} & \gamma \in \mathcal{M}_+(Z) \\ & \displaystyle\int_Z \phi_i(z)\, \mathrm{d}\gamma(z) = v_i \quad \forall i = 1, \ldots, m, \end{array}$$

and denote by $\mathcal{F}$ the feasible region of (4.4) and by $\mathrm{ext}(\mathcal{F})$ the set of extreme points of $\mathcal{F}$. Then, the following result can be found in (Yue et al., 2022).

Proposition 4.11 ((Yue et al., 2022, Prop. 1)). *Suppose that for all $\gamma \in \mathcal{F}$, at least one of the integrals $\int_X [\psi(z)]_+ \, \mathrm{d}\gamma(z)$ and $\int_X [-\psi(z)]_+ \, \mathrm{d}\gamma(z)$ is finite and that $\int_Z |\phi_i|(z)\, \mathrm{d}\gamma(z) < \infty$ for all $i = 1, \ldots, m$. If*

$$\sup \left\{ \int_Z \psi(z)\, \mathrm{d}\gamma(z) : \gamma \in \mathcal{F} \right\} \;=\; \sup \left\{ \int_Z \psi(z)\, \mathrm{d}\gamma(z) : \gamma \in \mathrm{ext}(\mathcal{F}) \right\}, \tag{4.5}$$

then it holds that

$$\sup \left\{ \int_Z \psi(z)\, \mathrm{d}\gamma(z) : \gamma \in \mathcal{F} \right\} \;=\; \sup \left\{ \int_Z \psi(z)\, \mathrm{d}\gamma : \gamma \in \mathcal{F} \cap \mathcal{D}_m(Z) \right\},$$

where $\mathcal{D}_m(Z)$ is the set of non-negative discrete measures supported on at most m points in Z. Furthermore, if $\mathcal{F} \subseteq \mathcal{P}(Z)$ and Z is Hausdorff, then the condition (4.5) is satisfied. This means that if ones add a constraint $\phi_{m+1} = \mathbb{1}_Z$ and $v_{m+1} = 1$ (so that $\mathcal{F} \subseteq \mathcal{P}(Z)$), then the condition (4.5) holds.

Proof. This is a direct consequence of (Pinelis, 2016, Coro. 5 and Prop. 6(v)). See also (Shapiro et al., 2021, Chap.2,Th. 60). □

4.3 Comparing distributions

In order to pursue our objective of seeing how robust solutions can be, we have to measure how nasty distributions can be. For this, we will properly define nastiness as the surprising nature or perplexity of a random variable which is well characterized by Shannon's entropy. This will then lead us to consider how noise affects functions. Finally, we will review some ways to compare distributions.

4.3.1 Information and Random Variables

The notion of the information brought by the outcome of a random variable has been introduced in the 1940's as the foundation of the field of information theory (see the foundational paper (Shannon, 1948)), which is not about *information* per se bet rather how transmissions and their incubent noise affect the amount of information (i.e., bits) that can be transmitted. The general idea is that if the outcome of a random variable is certain before observing it (e.g. $\mu(X = x) = 1$), the its observation is not informative. Similarly, knowing that a certain number will *not* be draw in a lottery is not very informative (as it is highly probable) while knowing that one number *will* be draw is very informative.

Shannon's characterization[17] of perplexity/self-information was chosen so that to meet several axioms:

- An event with probability 100% is perfectly unsurprising and yields no information
- The less probable an event is, the more surprising it is and the more information it yields
- If two independent events are measured separately, the total amount of information is the sum of the self-informations of the individual events

This means that for two independent event A and B, we seek a function h such that

- $h(A) = 0$ if $\mu(A) = 1$ and $h(A) > 0$ if $\mu(A) < 1$
- $h(A) = g(\mu(A))$ with g monotonically decreasing in $[0, 1]$
- $h(A \cap B) = h(A) + h(B)$

then the last two axioms imply that $g(\mu(A \cap B)) = g(\mu(A) \cdot \mu(B)) = g(\mu(A)) + g(\mu(B))$. This means that g has to verify Cauchy's logarithmic equation (i.e., $g(x \cdot y) = g(x) + g(y)$ + monotonicity) and for this the only solution is the logarithmic function (up to some scalar/base[18]).

[17] Later on, several other charaterizations were provided, see the wikipedia page of Entropy (Information theory)

[18] Different choices of base correspond to different units of information: base 2, the unit is the shannon (symbol Sh), often called a 'bit'; when base e, the unit is the natural unit of information (symbol nat); and base 10, the unit is the hartley (symbol Hart). We will stick with the natural log in order to streamline the presentation.

> *Solution of Cauchy's functional inequality (classical + logarithmic).*
>
> *Cauchy's functional inequality over rationals* Let $f : \mathbb{Q} \to \mathbb{Q}$ satisfy
>
> $$f(x + y) = f(x) + f(y) \quad \text{for all } x, y \in \mathbb{Q}.$$
>
> We aim to show that $f(x) = cx$ for some $c \in \mathbb{Q}$.
> Set $y = 0$:
>
> $$f(x + 0) = f(x) + f(0),$$
>
> which gives $f(0) = 0$.
> Now, for $x \in \mathbb{Q}$, setting $y = -x$:
>
> $$f(x + (-x)) = f(0) \implies f(x) + f(-x) = 0.$$
>
> Thus, $f(-x) = -f(x)$.
> For $n \in \mathbb{N}$, by induction, we have $f(nx) = nf(x)$ and $f(-nx) = -f(nx) = -nf(x)$.
> Let $x \in \mathbb{Q}$ and $x = \frac{p}{q}$ with $p, q \in \mathbb{Z}$, $q > 0$. Using additivity and scaling:
>
> $$f\left(\frac{p}{q}\right) = f\left(\frac{1}{q} + \cdots + \frac{1}{q}\right) = pf(\frac{1}{q}) = \frac{p}{q} f(1).$$
>
> Hence, the solution to Cauchy's functional equation over $\mathbb{Q}$ is $f(x) = cx$, where $c = f(1) \in \mathbb{Q}$.

Extension to Real Numbers for Monotonic Functions On $\mathbb{R}$, the situation is way more complex but some additional assumptions suffice to get back to the rational case. Suppose f is monotonic. Without loss of generality, assume f is non-decreasing. Since f is monotonic, it is continuous almost everywhere.
For $x \in \mathbb{R}$, consider a sequence $(q_n) \subset \mathbb{Q}$ such that $q_n \to x$. By monotonicity

$$f(q_n) = cq_n \quad \text{and hence} \quad \lim_{n\to\infty} f(q_n) = c \lim_{n\to\infty} q_n = cx.$$

Since f is monotonic, $f(x) = cx$ for all $x \in \mathbb{R}$.
Logarithmic case For any $x, y > 0$, writing $x = \exp(u)$ and $y = \exp(v)$, we have $g(x \cdot y) = g(\exp(u+v)) = g(x) + g(y) = g(\exp(u)) + g(\exp(v))$ and so Cauchy's equation applied to $f \equiv g \circ \exp$ gives that $g \circ \exp(u) = cu$ for all $u \in \mathbb{R}$ and thus $g(x) = c\log(x)$ for all $x \in \mathbb{R}_+^*$. □

This means that our measure of surprise for probability μ, sometimes called *self-information* is the function $h(A) = -\log(\mu(A))$ for any event A.

From this axiomatic definition, Shannon introduced the concept of entropy.[19] The *entropy* of a random variable X is naturally defined as the expected self-information. Here, we see an issue arising between the discrete and continuous case due to the definition of the probability of an event...

[19] "My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more importantly, no one knows what entropy really is, so in a debate you will always have the advantage." See (Rioul, 2021)

4.3.2 Entropy (the discrete case)

Let X follow some probability μ on a finite set $\mathcal{X}$ (or equivalently μ is an atomic distribution on $\mathcal{X}$).

Definition 4.12 (Entropy)**.** Denote by $p : \mathcal{X} \to [0,1]$ the discrete probabilities of the elements of $\mathcal{X}$ ($p(x) = \mu(X = x)$). Then, the *entropy* H is defined as

$$H(X) = \mathbb{E}[-\log p(X)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

with the convention $0\log 0 = 0$. We readily notice that H depends on X only through p (and not $\mathcal{X}$), so the abuse of notation $H(p)$ will be often used.

The notation H comes from Boltzmann's quantity H that was introduced in the 1870's in the context of statistical mechanics and thermodynamics and shares a similar formulation (see also the notion of Gibbs entropy).

The following properties are easily derived.

Lemma 4.13. *The entropy of a random variable verifies the following properties*

(a) $H(X) \geq 0$

(b) $H(X) \leq \log|\mathcal{X}|$ *where* $|\mathcal{X}|$ *is the number of elements in* $\mathcal{X}$*, with equality if and only if* X *has a uniform distribution over* $\mathcal{X}$

(c) $H(p)$ *is concave in* p

(d) If X *and* X' *are iid., then* $\mathbb{P}(X = X') \geq \exp(-H(X))$

Proof. Left as an exercise. A useful trick to recall is that $\log(x) \leq x - 1$ for all $x > 0$ with equality if and only if $x = 1$. For the probability of equality, use that $\exp\mathbb{E}\log(U) \leq \mathbb{E}\exp\log(U) = \mathbb{E}U$ for any rv U valued in $(0,1]$. □

Example 4.14 (Bernoulli variable). If $X \sim \mathcal{B}(q)$ then $H(X) = H(q) = -q\log(q) - (1-q)\log(1-q)$.

Since entropy is linked to information theory, it is often to consider two random variables and, by extending our notations, to define the *joint entropy* $H((X,Y)) = -\sum_{x,y\in\mathcal{X}} p(x,y)\log p(x,y)$ and the *conditional entropy* $H(Y|X) = -\sum_{x,y\in\mathcal{X}} p(x,y)\log p(y|x)$.

Then, the mutual information between X and Y is defined as the reduction of uncertainty of X due to the knowledge of Y.

Definition 4.15 (Mutual information)**.** The *mutual information* I between X and Y is defined as

$$I(X;Y) = \sum_{x,y\in\mathcal{X}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

with the convention $0\log 0 = 0$.

Then, we have the following properties.

Lemma 4.16. *The mutual information of a couple of random variables verifies the following properties*

(a) $I(X;Y) = I(Y;X)$

(b) $I(X;X) = H(X)$

(c) $I(X;Y) = H(X) - H(X|Y)$

(d) $I(X;Y) \geq 0$ *with equality if and only if* X *and* Y *are independent*

(e) If $X \to Y \to Z$ *(*X, Y, Z *form a Markov chain), then* $I(X;Y) \geq I(X;Z)$*; in particular,* $I(X;Y) \geq I(X;g(Y))$

Proof. Left as an exercise (For the non negativity, use Jensen's inequality). □

Example 4.17 (Horse races). Suppose that you bet on a race of m horses. You invest a fraction b_i of your money on horse i, which has probability p_i to win and return rate (or odd) of o_i. Thus, if i wins, your wealth grow by $S_i = b_i o_i$. The exponential rate of a horse race is $W(b,p) = \mathbb{E}(\log S) = \sum_{i=1}^m p_k \log(b_k o_k)$. Show that the optimal rate $W^\star(p)$ is obtained by taking $b = p$ and that $W^\star(p) = \sum_i p_i \log(o_i) - H(p)$. Furthermore, if the return rate is uniform $o_i = m$, then $W^\star(p) = \log(m) - H(p)$. Conclude.

4.3.3 Differential Entropy (the density case)

Let X follow some continuous probability distribution μ on set $\mathcal{X}$.

Definition 4.18 (Differential Entropy)**.** Denote by p the density of μ ($\mu(\mathrm{d}x) = \mathrm{d}\mu(x) = p(x)\mathrm{d}x$). Then, the *differential entropy* H is defined as

$$h(X) = \mathbb{E}[-\log p(X)] = -\int p(x)\log p(x)\,\mathrm{d}x$$

with the convention $0\log 0 = 0$, and provided that the integral exists. As in the discrete case, h depends on X only through p (and not $\mathcal{X}$), so the abuse of notation $H(p)$ will be often used.

Differential entropy is not invariant under a change of variables and can become negative. In addition, it is not even dimensionally correct. Since $h(X)$ would be dimensionless and $p(x)$ must have units of $\frac{1}{dx}$, this means that the argument to the logarithm is not dimensionless as required.

Example 4.19. What the entropy of a uniform random variable on $[0, a]$? For a Laplace rv? For a Gaussian rv? Compare for a fixed variance.

As previously, we can nevertheless ask which density maximizes the differential entropy.

Lemma 4.20. *Let us denote by $\mathcal{Q}$ the set of probability densities q on X such that $\int q(x) r_i(x)\,\mathrm{d}x = \alpha_i$ for $i = 1, .., m$ where the (r_i) are measurable functions and the (α_i) are real numbers. Then, the probability density $q^\star$ that maximizes $h(q)$ over $\mathcal{Q}$ is uniquely defined as $q(x) = \exp(\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x))$ for some $\lambda_0, .., \lambda_m$.*

Proof. See Exercise 4.3. □

Example 4.21 (Optimality of the Gaussian). The distribution on $\mathbb{R}$ with zero mean and variance σ^2 that has the largest entropy is the Gaussian distribution, attaining an entropy of $\log(2\pi\sigma^2)/2$.

Most properties on entropy fall but the ones on mutual information are preserved! Hence, wea can keep in mind that *comparing* random variables is more natural in an information theoretic perspective. This is the topic of the following section.

4.3.4 Kullback-Liebler Divergence

Bridging together the discrete and continuous cases, the Kullback-Liebler divergence is the relative entropy from the second measure to the first. We now drop the random variable dependence to a distribution dependence.

Definition 4.22 (Kullback-Liebler Divergence)**.** Let μ and ν be two probability measures on a measurable space X such that μ is absolutely continuous with respect to ν, then the relative entropy from Q to P is defined as

$$D_{\mathrm{KL}}(\mu||\nu) = \int_{x\in\mathrm{X}} \log\left(\frac{\mu(\mathrm{d}x)}{\nu(\mathrm{d}x)}\right)\mu(\mathrm{d}x)$$

where $\frac{\mu(\mathrm{d}x)}{\nu(\mathrm{d}x)}$ is the Radon-Nikodym derivative of μ with respect to ν.

We note that both the discrete and continuous case, $D_{\mathrm{KL}}(\mu||\nu) = H((\mu, \nu)) - H(\mu)$. We also have that the divergence between the joint and product of marginal distributions is the mutual information.

Lemma 4.23. *The Kullback-Liebler divergence of a couple of random variables is non-negative and null if and only if $\mu = \nu$ as measures.*

Proof. See Exercise 4.8. □

Finally, the Kullback-Liebler divergence is not a metric on the space of probability distributions. Indeed, it is not symmetric and does not satisfy the triangle inequality. However, it is a divergence (i.e., something that generalizes squared distances), and generates a topology in the space of distributions. A direct way to see this is through Pinsker's inequality.

Lemma 4.24. *Let μ and ν be two probability distributions on a measurable space X. Then,*

$$\|\mu - \nu\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(\mu||\nu)}$$

where $\|\mu - \nu\|_{TV} = \sup\{|\mu(A) - \nu(A)| \quad A \text{ is a measurable event}\}$ is the total variation distance *between μ and ν.*

Proof. See Exercise 4.10. □

Other related Distances & Divergences

- The family of Rényi divergences generalizes the Kullback-Liebler divergence
- The family of f-divergence is another way to generalize It

4.3.5 *Wasserstein distances*

4.3.6 *Integral probability metrics*

Take the class of linear functions on R^d:

$$\mathcal{F}_{\text{lin}} := \left\{ f(x) = \langle \theta, x \rangle \ : \theta, x \in \mathrm{R}^d, \ \|\theta\| = 1 \right\},$$

and consider the quantity

$$\begin{aligned} \delta(\mu, \nu) &= \sup_{f \in \mathcal{F}_{\text{lin}}} \left\{ \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right\} \\ &= \sup_{\theta \in \mathrm{R}^d, \, \|\theta\| = 1} \left\{ \int \langle \theta, x \rangle \, \mu(\mathrm{d}x) - \int \langle \theta, y \rangle \, \nu(\mathrm{d}y) \right\} \\ &= \|\mathbb{E}_\mu[X] - \mathbb{E}_\nu[Y]\| \, . \end{aligned}$$

In particular, $\delta(\mu, \nu) = 0$ if and only if μ and ν have the same mean. This is of course not sufficient to say that the two measures are the same so the above quantity does not define a distance between probability measures.

However, if the class $\mathcal{F}$ is large enough, for instance the 1-Lipchitz functions. Then, it can become a metric, called an integral probability metric.

Definition 4.25. A metric $\mathrm{d}(\cdot, \cdot)$ between two probability measures is called an *integral probability metric* (IPM) if it satisfies the properties of a metric and can be written in the form

$$\mathrm{d}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right| .$$

4.4 Distributional smoothing

Mirroring what we did before, we can "add noise" to optima as a way to create smoothness more easily.

Lemma 4.26. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function and $C \subset \mathbb{R}^n$, then*

$$\sup_{x \in C} f(x) = \sup_{\pi \in \mathcal{P}(C)} \int_C f(t) \mathrm{d}\pi(t)$$

where $\mathcal{P}(C)$ denotes the set of probability distributions over C

Proof. Left as an exercise □

Then the same principle can be applied for convex conjugates (Clason and Valkonen, 2020, Chap. 5) and the smoothness comes from (Clason and Valkonen, 2020, Chap. 7.1).

Theorem 4.27. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function and $C \subset \mathbb{R}^n$ compact, then*

$$\sup_{\pi \in \mathcal{P}(C)} \int_C f(t) \mathrm{d}\pi(t) - \Omega(\pi) := \Omega^*(f)$$

Thus, if Ω is μ-strongly convex, Ω^* is $1/\mu$-uniformly smooth.

4.4.1 Variational Characterization of the Kullback–Leibler Divergence

We follow (Agrawal and Horel, 2021) in this section dedicated to KL, but the paper considers general ϕ-divergences and not only probability measures.

Let $(\Omega, \mathcal{F})$ be a measurable space and let ν be a probability measure on Ω. For any probability measure μ on Ω, we recall that the Kullback–Leibler divergence is defined by

$$D_{\mathrm{KL}}(\mu \| \nu) = \begin{cases} \displaystyle\int_\Omega \log\left(\frac{d\mu}{d\nu}\right) d\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

For a measurable function $f : \Omega \to \mathbb{R}$ define the log-partition functional

$$\Lambda_\nu(f) := \log \int_\Omega e^f \, d\nu,$$

with the convention $\Lambda_\nu(f) = +\infty$ if the integral diverges.

Theorem 4.28 (Donsker–Varadhan formula)**.** *For every probability measure μ,*

$$D_{\mathrm{KL}}(\mu \| \nu) = \sup_f \left\{ \int_\Omega f \, d\mu - \log \int_\Omega e^f \, d\nu \right\},$$

where the supremum is taken over all measurable f such that $\int e^f \, d\nu < \infty$.

Proof. We prove the two inequalities.

Step 1: Lower bound. Assume $\mu \ll \nu$ and write $p = \frac{d\mu}{d\nu}$. For $M > 0$ define $f_M := \log(p \wedge M)$. Then

$$\int f_M \, d\mu - \log \int e^{f_M} \, d\nu = \int \log(p \wedge M) \, p \, d\nu - \log \int (p \wedge M) \, d\nu.$$

Since $\int (p \wedge M) \, d\nu \le 1$, the second term is nonnegative, hence

$$\int f_M \, d\mu - \log \int e^{f_M} \, d\nu \le \int \log(p \wedge M) \, p \, d\nu.$$

By monotone convergence,

$$\int \log(p \wedge M) \, p \, d\nu \uparrow \int p \log p \, d\nu = D_{\mathrm{KL}}(\mu \| \nu).$$

Thus the supremum is at least $D_{\mathrm{KL}}(\mu\|\nu)$.
If $\mu \not\ll \nu$, choose A with $\mu(A) > 0$, $\nu(A) = 0$ and let $f_n = n1_A$. Then

$$\int f_n \, d\mu = n\mu(A) \to \infty, \qquad \int e^{f_n} \, d\nu = 1,$$

so the supremum is $+\infty$, matching $D_{\mathrm{KL}}(\mu\|\nu)$.

Step 2: Upper bound. Assume $\mu \ll \nu$ with density p. For any measurable f,

$$\int f \, d\mu - \log \int e^f \, d\nu = \int (f - \log p)p \, d\nu + \int p \log p \, d\nu - \log \int e^f \, d\nu.$$

By Jensen's inequality applied to the probability measure $p \, d\nu$,

$$\int (f - \log p)p \, d\nu \le \log \int e^{f - \log p} p \, d\nu = \log \int e^f \, d\nu.$$

Rearranging gives

$$\int f \, d\mu - \log \int e^f \, d\nu \le \int p \log p \, d\nu = D_{\mathrm{KL}}(\mu\|\nu).$$

Taking the supremum over f yields the reverse inequality. □

Theorem 4.29. *Let $\Lambda_\nu(f) = \log \int e^f \, d\nu$. Then its convex conjugate is*

$$\Lambda_\nu^*(\mu) = D_{\mathrm{KL}}(\mu\|\nu).$$

Proof. By definition,

$$\Lambda_\nu^*(\mu) = \sup_f \left\{ \int f \, d\mu - \Lambda_\nu(f) \right\}.$$

This is exactly the variational formula of Theorem 4.28. □

4.4.2 Log-Sum-Exp as the Dual of the Kullback–Leibler Divergence

We now establish the dual representation of the log-sum-exp functional.

Theorem 4.30 (Log-sum-exp as convex conjugate of KL)**.** *Let ν be a probability measure on Ω. For any measurable function $f : \Omega \to \mathbb{R}$ such that $\int e^f \, d\nu < \infty$, one has*

$$\log \int_\Omega e^f \, d\nu = \sup_{\mu \in \mathcal{P}(\Omega)} \left\{ \int_\Omega f \, d\mu - D_{\mathrm{KL}}(\mu\|\nu) \right\}.$$

Proof. We prove the equality in two steps.

Step 1: Upper bound. Let $\mu \in \mathcal{P}(\Omega)$ be arbitrary. By the variational representation of the KL divergence (Theorem 4.28),

$$D_{\mathrm{KL}}(\mu\|\nu) = \sup_g \left\{ \int g \, d\mu - \log \int e^g \, d\nu \right\}.$$

In particular, choosing $g = f$ gives

$$D_{\mathrm{KL}}(\mu\|\nu) \ge \int f \, d\mu - \log \int e^f \, d\nu,$$

which rearranges to

$$\int f\,d\mu - D_{\mathrm{KL}}(\mu\|\nu) \le \log \int e^f\,d\nu.$$

Since this holds for all μ, we obtain

$$\sup_{\mu\in\mathcal{P}(\Omega)} \left\{ \int f\,d\mu - D_{\mathrm{KL}}(\mu\|\nu) \right\} \le \log \int e^f\,d\nu.$$

Step 2: Achieving equality. Define the probability measure μ_f by its Radon-Nikodym derivative with respect to ν as

$$\frac{d\mu_f}{d\nu} = \frac{e^f}{\int e^f\,d\nu}.$$

Then $\mu_f \ll \nu$ and

$$D_{\mathrm{KL}}(\mu_f\|\nu) = \int \log\left(\frac{e^f}{\int e^f\,d\nu}\right) d\mu_f = \int f\,d\mu_f - \log \int e^f\,d\nu.$$

Rearranging yields

$$\int f\,d\mu_f - D_{\mathrm{KL}}(\mu_f\|\nu) = \log \int e^f\,d\nu.$$

Thus the supremum is attained at μ_f, and equality holds.

□

4.4.3 Summary

The previous results establish the exact convex duality:

$$\boxed{\log \int e^f\,d\nu = \sup_{\mu\in\mathcal{P}(\Omega)} \left\{ \int f\,d\mu - D_{\mathrm{KL}}(\mu\|\nu) \right\}}$$

and, equivalently,

$$\boxed{D_{\mathrm{KL}}(\mu\|\nu) = \sup_f \left\{ \int f\,d\mu - \log \int e^f\,d\nu \right\}.}$$

These identities form a Legendre-Fenchel dual pair:

$$\log \int e^{(\cdot)}\,d\nu \quad \longleftrightarrow \quad D_{\mathrm{KL}}(\cdot\|\nu).$$

- $\log \int e^f\,d\nu$ is the cumulant generating functional, it is convex and lower semi-continuous.
- D_{KL} is its convex conjugate.
- The optimizer is the Gibbs measure

$$d\mu_f = \frac{e^f}{\int e^f\,d\nu}\,d\nu.$$

- This duality underlies large deviations, PAC-Bayes bounds, and variational inference.
- The log-sum-exp duality is only valid for *probability distributions* (see Exercise 4.13).

Exercises

In this whole part, we let $(\Omega, \mathcal{A})$ be a measurable space and let μ and ν be two probability measures on $(\Omega, \mathcal{A})$. Suppose that τ is a σ-finite measure on $(\Omega, \mathcal{A})$ satisfying $\mu \ll \tau$ and $\nu \ll \tau$. Define $p = \mathrm{d}\mu/\mathrm{d}\tau, q = \mathrm{d}\nu/\mathrm{d}\tau$. Observe that such a measure τ always exists since we can take, for example, $\tau = \mu + \nu$.

Exercise 4.1. In the same setting as Example 4.3, suppose that $F_n(x) := \mathbb{E}_{X\sim\mu_n}[f(x;X)] \to \mathbb{E}_{X\sim\mu}[f(x;X)] = F(x)$ for all x. Suppose in addition that the functions (F_n) are convex and that F has a unique minimizer. Do we have a convergence of the $\arg\min$s to the minimizer of F?

Exercise 4.2. In the setting of Remark 4.2, show that in the convex case, if $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$, then x_k converges almost surely to a minimizer of F.

Elements of Solution: Let $x^\star$ be a minimizer of F. Then,

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^\star\|^2 | \mathcal{F}_k\right] &= \mathbb{E}\left[\|x_k - x^\star - \gamma_k \nabla_x f(x_k; X_{k+1})\|^2 | \mathcal{F}_k\right] \\
&= \|x_k - x^\star\|^2 + \gamma_k^2 \mathbb{E}\left[\|\nabla_x f(x_k; X_{k+1})\|^2 | \mathcal{F}_k\right] - 2\gamma_k \langle x_k - x^\star, \nabla_x F(x_k)\rangle \\
&= \|x_k - x^\star\|^2 + \gamma_k^2 \|\nabla_x F(x_k) - \nabla_x F(x^\star)\|^2 - 2\gamma_k \langle x_k - x^\star, \nabla_x F(x_k) - \nabla_x F(x^\star)\rangle \\
&\quad + \gamma_k^2 \mathbb{E}\left[\|\nabla_x f(x_k; X_{k+1}) - \mathbb{E}[\nabla_x f(x_k; X_{k+1}) | \mathcal{F}_k]\|^2 | \mathcal{F}_k\right] \\
&\leq \|x_k - x^\star\|^2 + \left(\gamma_k^2 - \frac{\gamma_k}{L}\right) \|\nabla_x F(x_k) - \nabla_x F(x^\star)\|^2 \\
&\quad + \gamma_k^2 \mathbb{E}\left[\|\nabla_x f(x_k; X_{k+1}) - \mathbb{E}[\nabla_x f(x_k; X_{k+1}) | \mathcal{F}_k]\|^2 | \mathcal{F}_k\right] \\
&\leq (1 + \gamma_k^2)\|x_k - x^\star\|^2 - \frac{\gamma_k}{L}\|\nabla_x F(x_k) - \nabla_x F(x^\star)\|^2 + \gamma_k^2 \sigma^2
\end{aligned}$$

where in the first inequality we use the smoothness *and the convexity* (see (Bubeck et al., 2015, Lem. 3.5)). We are now in position to use Robbins-Siegmund theorem to show that $\|x_{k+1} - x^\star\|^2$ converges almost surely. With some additional technicalities, we show the claimed result. □

Exercise 4.3. Prove Lemma 4.20

Exercise 4.4. Prove the claim of Example 4.21.

Exercise 4.5. Let $a \in \mathbb{N}$. What is the maximum (discrete) entropy distribution on $\{0, 1, .., a\}$? What is the maximum (differential) entropy distribution on $[0, a]$?

Exercise 4.6. If X is compact (say an interval) and we consider the discretized version of X, called X^Δ, where Δ is the discretization step. Show that $H(X^\Delta) + \log(\Delta) \to_{\Delta\to 0} h(X)$ and thus that a n-bits discretization of has an entropy of approximately $h(X) + cn$ where c is a constant.

Exercise 4.7. What is the maximal entropy discrete distribution with a prescribed mean on an infinite set? How does this relate to the questions above? See the related Wikipedia page.

Exercise 4.8 (Proof of Lemma 4.23)**.** Prove that

$$D_{KL}(\mu\|\nu) = 0 \quad \text{if and only if} \quad \mu = \nu.$$

Elements of Solution: If $\mu \not\ll \nu$, $D_{KL}(\mu\|\nu) = +\infty$ and $\mu \neq \nu$ so we focus on the case where $\mu \ll \nu$. Since $\mu \ll \nu$, by the Radon-Nikodym theorem, there exists a measurable function $f = \frac{d\mu}{d\nu} : \Omega \to [0, \infty)$ such that

$$\mu(A) = \int_A f(x) d\nu(x)$$

for all $A \in \mathcal{F}$. By definition, the KL divergence is given by

$$D_{KL}(\mu\|\nu) = \int_\Omega f(x) \log\left(f(x)\right) d\nu(x).$$

(i) If $\mu = \nu$ then $D_{KL}(\mu\|\nu) = 0$: If $\mu = \nu$, then for ν-almost every $x \in \Omega$ we have

$$f(x) = \frac{d\mu}{d\nu}(x) = 1.$$

Hence,

$$D_{KL}(\mu\|\nu) = \int_\Omega 1 \cdot \log(1) d\nu(x) = \int_\Omega 0 \, d\nu(x) = 0.$$

(ii) If $D_{KL}(\mu\|\nu) = 0$ then $\mu = \nu$: Assume that

$$D_{KL}(\mu\|\nu) = \int_\Omega f(x) \log\left(f(x)\right) d\nu(x) = 0.$$

For any $t \geq 0$, the function

$$\varphi(t) = t \log t - t + 1$$

satisfies $\varphi(t) \geq 0$ with equality if and only if $t = 1$. As $t \log t = \varphi(t) + t - 1$, we can rewrite the KL divergence as

$$D_{KL}(\mu\|\nu) = \int_\Omega \left[\varphi(f(x)) + f(x) - 1\right] d\nu(x).$$

Since

$$\int_\Omega f(x) d\nu(x) = \mu(\Omega) = 1, \quad \int_\Omega 1 d\nu(x) = \nu(\Omega) = 1,$$

we obtain

$$0 = D_{KL}(\mu\|\nu) = \int_\Omega \varphi(f(x))d\nu(x).$$

Since $\varphi(f(x)) \geq 0$ for all x and the integral is zero, it follows that

$$\varphi(f(x)) = 0 \quad \text{for } \nu\text{-almost every } x.$$

By the characterization of φ, we deduce that $f(x) = 1$ for ν-almost every x. Therefore,

$$\frac{d\mu}{d\nu}(x) = 1 \quad \text{for } \nu\text{-almost every } x,$$

which implies that

$$\mu(A) = \int_A 1 d\nu(x) = \nu(A)$$

for all $A \in \mathcal{F}$. That is, $\mu = \nu$.

Thus, we have shown that $D_{KL}(\mu\|\nu) = 0$ if and only if $\mu = \nu$. □

Exercise 4.9 (Total Variation)**.** The total variation distance between μ and ν is defined as follows:

$$\|\mu - \nu\|_{TV} = \sup_{A\in\mathcal{A}} |\mu(A) - \nu(A)| = \sup_{A\in\mathcal{A}} \left|\int_A (p-q)d\tau\right|.$$

1. Show Scheffé's theorem stating that

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\int |p-q|\mathrm{d}\tau = 1 - \int \min(\mathrm{d}\mu, \mathrm{d}\nu) = \int \max(\mathrm{d}\mu, \mathrm{d}\nu) - 1$$

2. Deduce that $0 \leq \|\mu - \nu\|_{TV} \leq 1$ and that the total variation satisfies the axioms of distance.

Elements of Solution: Let $A_0 = \{x \in \Omega : q(x) \geq p(x)\}$. Then

$$\begin{aligned}\int |p-q|\mathrm{d}\tau &= \int_{A_0} (q-p)\mathrm{d}\tau + \int_{A_0^c} (p-q)\mathrm{d}\tau \\ &= 2\int_{A_0} (q-p)\mathrm{d}\tau\end{aligned}$$

where A_0^c is the complement of A_0 and we use that $\int_{A_0^c} p\mathrm{d}\tau = 1 - \int_{A_0} p\mathrm{d}\tau$. We also

have

$$\begin{aligned}\int |p-q|\mathrm{d}\tau &= \int_{A_0}(q-p)\mathrm{d}\tau + \int_{A_0^c}(p-q)\mathrm{d}\tau\\ &= \int_{A_0}(q-\min(p,q))\mathrm{d}\tau + \int_{A_0^c}(p-\min(p,q))\mathrm{d}\tau\\ &= \int_{A_0} q\mathrm{d}\tau + \int_{A_0^c} p\mathrm{d}\tau - \int \min(p,q)\mathrm{d}\tau\\ &= \int_{A_0}(q-p)\mathrm{d}\tau + 1 - \int \min(p,q)\mathrm{d}\tau\\ &= \frac{1}{2}\int |p-q|\mathrm{d}\tau + 1 - \int \min(p,q)\mathrm{d}\tau\end{aligned}$$

and thus $\frac{1}{2}\int |p-q|\mathrm{d}\tau = 1 - \int \min(p,q)\mathrm{d}\tau$ and similarly, $\frac{1}{2}\int |p-q|\mathrm{d}\tau = \int \max(p,q)\mathrm{d}\tau - 1$. Hence, we have

$$\|\mu-\nu\|_{TV} \geq \nu(A_0) - \mu(A_0) = \frac{1}{2}\int |p-q|\mathrm{d}\tau = 1 - \int \min(p,q)\mathrm{d}\tau$$

On the other hand, for all $A \in \mathcal{A}$,

$$\begin{aligned}\left|\int_A (q-p)\mathrm{d}\tau\right| &= \left|\int_{A\cap A_0}(q-p)\mathrm{d}\tau + \int_{A\cap A_0^c}(q-p)\mathrm{d}\tau\right|\\ &\leq \max\left\{\int_{A_0}(q-p)\mathrm{d}\tau, \int_{A_0^c}(p-q)\mathrm{d}\tau\right\} = \frac{1}{2}\int |p-q|\mathrm{d}\tau\end{aligned}$$

Then

$$\|\mu-\nu\|_{TV} = \nu(A_0) - \mu(A_0)$$

implying the required result. □

Exercise 4.10 (Proof of Pinsker's Inequality)**.** Suppose that $\mu \ll \nu$. Show that

$$\|\mu-\nu\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(\mu\|\nu)}.$$

Elements of Solution: Assume that $\mu \ll \nu$ and denote

$$f = \frac{d\mu}{d\nu}.$$

Then,

$$D_{KL}(\mu\|\nu) = \int f \log f \, d\nu,$$

and

$$\|\mu-\nu\|_{TV} = \frac{1}{2}\int |f-1|\, d\nu.$$

Let us consider the function

$$\phi(t) = t\log t - t + 1,$$

for which one can show that

$$\left(\frac{4}{3}+\frac{2}{3}t\right)\phi(t) \geq (t-1)^2$$

holds for all $t \geq 0$.
Integrate the inequality with respect to $d\nu$:

$$\begin{aligned}
\|\mu-\nu\|_{TV} = \frac{1}{2}\int |f-1|\,d\nu &\leq \frac{1}{2}\int \sqrt{\left(\frac{4}{3}+\frac{2}{3}f\right)\left(f\log f - f + 1\right)}\,d\nu \\
&\leq \frac{1}{2}\sqrt{\int\left(\frac{4}{3}+\frac{2}{3}f\right)d\nu}\sqrt{\int\left(f\log f - f+1\right)d\nu} \\
&= \frac{1}{2}\sqrt{\frac{4}{3}\int d\nu + \frac{2}{3}\int d\mu}\sqrt{\int\left(f\log f - f+1\right)d\nu} \\
&= \sqrt{\frac{1}{2}\int\left(f\log f - f+1\right)d\nu} \\
&= \sqrt{\frac{1}{2}D_{KL}(\mu\|\nu)}
\end{aligned}$$

where we used the above inequality and Cauchy-Schwarz. This completes the proof of Pinsker's inequality □

Exercise 4.11 (Pinsker's Inequality for Discrete Distributions)**.** Let $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ be two probability distributions on the finite set $\{1, 2, \ldots, n\}$. Prove that

$$\|p-q\|_1 \leq \sqrt{2\,D_{KL}(p\|q)}.$$

Elements of Solution: In the discrete case, the total variation distance is given by

$$\|p-q\|_{TV} = \frac{1}{2}\|p-q\|_1.$$

By applying the result of Exercise 4.10 to the discrete measures p and q (taking the Radon–Nikodym derivative to be $f(i) = p_i/q_i$), we have

$$\|p-q\|_{TV} \leq \sqrt{\frac{1}{2}D_{KL}(p\|q)}.$$

Multiplying both sides by 2 yields

$$\|p-q\|_1 \leq \sqrt{2\,D_{KL}(p\|q)}.$$

This completes the proof for the discrete setting. □

Exercise 4.12 (Total Variation and IPM)**.**

The *total variation* distance between μ and ν is defined by

$$d_{TV}(\mu,\nu) = \sup_{A\in\mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)|.$$

Show that TV is an IPM (see Definition 4.25) for $\mathcal{F} = \{f : \|f\|_\infty \le 1\}$.

Elements of Solution: For any measurable set A, define the indicator function $f_A(x) = 1_A(x)$. Since $\|f_A\|_\infty \le 1$, we have

$$\mu(A) - \nu(A) = \int_{\mathcal{X}} f_A(x)\, d\mu(x) - \int_{\mathcal{X}} f_A(x)\, d\nu(x).$$

Taking the supremum over all measurable sets A, it follows that

$$\sup_{A\in\mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)| \le \sup_{\|f\|_\infty\le 1} \left| \int f\, d\mu - \int f\, d\nu \right|$$

but, defining $\tilde{f}_A(x) = 1_A(x) - 1_{A^c}(x)$, we still have $\|\tilde{f}_A\|_\infty \le 1$ but $\int_{\mathcal{X}} \tilde{f}_A(x)\, d\mu(x) = \mu(A) - (1-\mu(A))$. Hence, $\int_{\mathcal{X}} \tilde{f}_A(x)\, d\mu(x) - \int_{\mathcal{X}} \tilde{f}_A(x)\, d\nu(x) = 2(\mu(A) - \nu(A))$.
For the reverse inequality, let A^* be a measurable set such that

$$d_{TV}(\mu,\nu) = |\mu(A^*) - \nu(A^*)|.$$

Define

$$f(x) = \begin{cases} 1, & x \in A^*, \\ -1, & x \notin A^*. \end{cases}$$

Clearly, $\|f\|_\infty = 1$, and then

$$\int f\, d\mu - \int f\, d\nu = [\mu(A^*) - \mu(A^{*c})] - [\nu(A^*) - \nu(A^{*c})].$$

Since $\mu(A^*) + \mu(A^{*c}) = \nu(A^*) + \nu(A^{*c}) = 1$, one can verify that

$$\int f\, d\mu - \int f\, d\nu = 2\Big[\mu(A^*) - \nu(A^*)\Big] = 2\, d_{TV}(\mu,\nu).$$

Thus,

$$\sup_{\|f\|_\infty\le 1} \left| \int f\, d\mu - \int f\, d\nu \right| \ge 2\, d_{TV}(\mu,\nu).$$

Combining the two inequalities, we obtain

$$\sup_{\|f\|_\infty\le 1} \left| \int f\, d\mu - \int f\, d\nu \right| = 2\, d_{TV}(\mu,\nu),$$

or equivalently,

$$d_{TV}(\mu,\nu) = \frac{1}{2} \sup_{\|f\|_\infty\le 1} \left| \int f\, d\mu - \int f\, d\nu \right|.$$

This shows that TV is an IPM with function class $\{f : \|f\|_\infty \le 1\}$. □

Exercise 4.13. Let $(X, \mathcal{B})$ be a measurable space and let ν be a finite positive measure. Let $\mathcal{M}_+(X)$ denote the space of finite nonnegative measures on X.
For $\mu \in \mathcal{M}_+(X)$, define the (extended) Kullback–Leibler divergence

$$\mathrm{KL}(\mu\|\nu) = \begin{cases} \displaystyle\int \log\left(\frac{d\mu}{d\nu}\right) d\mu - \mu(X) + \nu(X), & \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

This definition coincides with the usual KL divergence when μ, ν are probability measures.
Show that for any measurable function $f : X \to \mathbb{R}$ such that $\int e^f \, d\nu < \infty$, one has

$$\sup_{\mu \in \mathcal{M}_+(X)} \left\{ \int f \, d\mu - \mathrm{KL}(\mu\|\nu) \right\} = \int (e^f - 1) \, d\nu \,.$$

Elements of Solution:
We split the proof into two steps.

Step 1: Upper bound.
Let $\mu \ll \nu$ with density $p = \frac{d\mu}{d\nu}$. Then

$$\int f \, d\mu - \mathrm{KL}(\mu\|\nu) = \int \big(fp - p \log p + p - 1 \big) \, d\nu.$$

Using the elementary inequality

$$ab - a \log a \le e^b - 1 \quad \text{for all } a \ge 0,\ b \in \mathbb{R},$$

we obtain

$$fp - p \log p \le e^f - 1.$$

Hence

$$\int f \, d\mu - \mathrm{KL}(\mu\|\nu) \le \int (e^f - 1) \, d\nu.$$

Since this holds for all $\mu \in \mathcal{M}_+(X)$,

$$\sup_{\mu} \left\{ \int f \, d\mu - \mathrm{KL}(\mu\|\nu) \right\} \le \int (e^f - 1) \, d\nu.$$

Step 2: Optimality.
Define the measure μ_f by

$$d\mu_f = e^f \, d\nu.$$

Then $\mu_f \ll \nu$ and

$$\mathrm{KL}(\mu_f\|\nu) = \int f e^f \, d\nu - \int e^f \, d\nu + \nu(X).$$

Therefore,

$$\int f \, d\mu_f - \mathrm{KL}(\mu_f\|\nu) = \int (e^f - 1) \, d\nu.$$

This achieves the upper bound, proving the claim.

Relation to the Probability Case

If μ, ν are probability measures, then

$$\int (e^f - 1)\, d\nu = \log \int e^f \, d\nu \quad \text{after normalization}.$$

Indeed, we had

$$\sup_{\mu \in \mathcal{M}_+(X)} \left\{ \int f \, d\mu - \mathrm{KL}(\mu \| \nu) \right\} = \int (e^f - 1)\, d\nu.$$

We now impose the constraint $\mu(X) = 1$, i.e. $\mu \in \mathcal{P}(X)$ by using a KKT multiplier $\lambda \in \mathbb{R}$:

$$\sup_{\mu \geq 0} \left\{ \int f \, d\mu - \mathrm{KL}(\mu \| \nu) - \lambda(\mu(X) - 1) \right\}.$$

Rewriting,

$$= \sup_{\mu \geq 0} \left\{ \int (f - \lambda)\, d\mu - \mathrm{KL}(\mu \| \nu) \right\} + \lambda.$$

Using what we had previously with f replaced by $f - \lambda$, we obtain

$$\sup_{\mu \geq 0} \left\{ \int (f - \lambda)\, d\mu - \mathrm{KL}(\mu \| \nu) \right\} = \int (e^{f-\lambda} - 1)\, d\nu.$$

Hence,

$$\sup_{\mu \in \mathcal{P}(X)} \left\{ \int f \, d\mu - D_{\mathrm{KL}}(\mu \| \nu) \right\} = \int (e^{f-\lambda} - 1)\, d\nu + \lambda.$$

Now, to find the value of the optimal multiplier, define

$$\Phi(\lambda) = \int e^{f-\lambda} \, d\nu + \lambda - 1.$$

Then

$$\Phi'(\lambda) = -\int e^{f-\lambda} \, d\nu + 1.$$

The minimum is attained when

$$\int e^{f-\lambda} \, d\nu = 1 \quad \Longleftrightarrow \quad \lambda = \log \int e^f \, d\nu.$$

Substituting back yields

$$\sup_{\mu \in \mathcal{P}(X)} \left\{ \int f \, d\mu - D_{\mathrm{KL}}(\mu \| \nu) \right\} = \log \int e^f \, d\nu.$$

□

Exercise 4.14 (Application in Distributionally Robust Optimization)**.** Suppose that the true probability distribution μ is unknown, but it is known to lie in a Kullback-Leibler

(KL) divergence ball around a nominal distribution ν:

$$\mathcal{U} = \{\mu : D_{KL}(\mu\|\nu) \leq \delta\}.$$

Show that for any measurable event A,

$$\sup_{\mu\in\mathcal{U}} |\mu(A) - \nu(A)| \leq \sqrt{\frac{\delta}{2}}.$$

Elements of Solution: By Pinsker's inequality (Exercise 4.10), for any $\mu \in \mathcal{U}$ we have

$$\|\mu - \nu\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(\mu\|\nu)} \leq \sqrt{\frac{\delta}{2}}.$$

Recall that for any measurable set A,

$$|\mu(A) - \nu(A)| \leq \|\mu - \nu\|_{TV}.$$

Thus,

$$\sup_{\mu\in\mathcal{U}} |\mu(A) - \nu(A)| \leq \sqrt{\frac{\delta}{2}},$$

which is the desired bound. □

Exercise 4.15 (Reduction to the Binary Case)**.** Show that for a fixed total variation distance, the Kullback-Leibler (KL) divergence is maximized by a two-point (binary) distribution.
Hint: prove that since KL divergence is a convex function of the density ratio and that the worst-case scenario occurs when the ratio takes on only two distinct values.

Exercise 4.16 (Alternative Pinsker Bound: Bretagnolle-Huber Inequality)**.** Let $\mu \ll \nu$. We want to prove that

$$\|\mu - \nu\|_{TV} \leq \sqrt{1 - \exp\big(-D_{KL}(\mu\|\nu)\big)} \leq 1 - \frac{1}{2}\exp\big(-D_{KL}(\mu\|\nu)\big).$$

1. Prove the following inequality:

$$1 - \|\mu - \nu\|_{TV}^2 \geq \left(\int \sqrt{pq}\,\mathrm{d}\tau\right)^2.$$

2. Use that $(\cdot)^2 = \exp(2\log(\cdot))$ to prove the following inequality:

$$\left(\int \sqrt{pq}\,\mathrm{d}\tau\right)^2 \geq \exp\left(2\int_{pq>0} p \log\sqrt{\frac{q}{p}}\,\mathrm{d}\tau\right) = \exp(-D_{KL}(\mu\|\nu))$$

Elements of Solution:

$$\begin{aligned}1 - \|\mu - \nu\|_{TV}^2 &= (1 - \|\mu - \nu\|_{TV})(1 + \|\mu - \nu\|_{TV}) \\ &= \int \min(\mathrm{d}\mu, \mathrm{d}\nu) \int \max(\mathrm{d}\mu, \mathrm{d}\nu) \\ &\geq \left(\int \sqrt{\min(\mathrm{d}\mu, \mathrm{d}\nu) \max(\mathrm{d}\mu, \mathrm{d}\nu)} \right)^2 \\ &= \left(\int \sqrt{pq} \mathrm{d}\tau \right)^2 .\end{aligned}$$

Writing $(\cdot)^2 = \exp(2\log(\cdot))$ and using Jensen's inequality we get:

$$\begin{aligned}\left(\int \sqrt{pq} \mathrm{d}\tau \right)^2 &= \exp\left(2 \log \int_{pq>0} \sqrt{pq} \mathrm{d}\tau \right) = \exp\left(2 \log \int_{pq>0} p \sqrt{\frac{q}{p}} \mathrm{d}\tau \right) \\ &\geq \exp\left(2 \int_{pq>0} p \log \sqrt{\frac{q}{p}} \mathrm{d}\tau \right) = \exp(-D_{KL}(\mu \| \nu))\end{aligned}$$

□

Exercise 4.17 (Application to Hypothesis Testing)**.** Consider a binary hypothesis testing problem between $H_0 : X \sim \nu$ and $H_1 : X \sim \mu$. Let ϕ be any test function with Type I error $\alpha = \nu(\phi(X) = 1)$ and Type II error $\beta = \mu(\phi(X) = 0)$. Prove that

$$\alpha + \beta \geq \frac{1}{2} \exp\big(-D_{KL}(\mu \| \nu)\big).$$

Hint: As an intermediate point, show that for any measurable set A,

$$\mu(A) + \nu(A^c) \geq \frac{1}{2} \exp\big(-D_{KL}(\mu \| \nu)\big).$$

Elements of Solution: For any measurable set A, $\|\mu - \nu\|_{TV} = \sup_{A' \in \mathcal{A}} |\mu(A') - \nu(A')| \geq \mu(A) - \nu(A) = 1 - (\mu(A^c) + \nu(A))$. Using Exercise 4.16, we get that

$$\begin{aligned}&1 - (\mu(A^c) + \nu(A)) \leq 1 - \frac{1}{2} \exp\big(-D_{KL}(\mu \| \nu)\big) \\ \Leftrightarrow &\mu(A^c) + \nu(A) \geq \frac{1}{2} \exp\big(-D_{KL}(\mu \| \nu)\big)\end{aligned}$$

Setting $A^c = \{x : \phi(x) = 1\}$,

$$\alpha = \nu(A), \quad \beta = \mu(A^c).$$

Thus,

$$\alpha + \beta \geq \frac{1}{2} \exp\big(-D_{KL}(\mu \| \nu)\big),$$

providing a lower bound on the error sum, showing that no test can have both errors arbitrarily small when D_{KL} is small. □

Exercise 4.18 (Dual Representation of ϕ-Divergences from (Kuhn et al., 2024, Prop. 2.6))**.** An entropy function $\phi : \mathrm{R} \to \overline{\mathrm{R}}$ is a lower semicontinuous convex function with $\phi(1) = 0$ and $\phi(s) = +\infty$ for all $s < 0$. This gives a set of ϕ-divergences

as

$$D_\phi(\mu, \hat{\mu}) = \int_{\mathrm{X}} \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z) \cdot \phi\left(\frac{\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(z)}{\frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z)}\right) \mathrm{d}\rho(z) \ .$$

Show that we have

$$D_\phi(\mu, \hat{\mu}) = \sup_{f \in \mathcal{F}} \int_{\mathrm{X}} f(z)\, \mathrm{d}\mu(z) - \int_{\mathrm{X}} \phi^*(f(z))\, \mathrm{d}\hat{\mu}(z),$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f : \mathrm{X} \to \mathrm{dom}(\phi^*)$.

Elements of Solution: As the entropy function $\phi(s)$ is proper, convex and lower semi-continuous on R and as $0\phi(s/0)$ is interpreted as the recession function $\phi^\infty(s)$, the perspective function $\phi^\pi(s,t) = t\phi(s/t)$ is proper, convex and lower semicontinuous on $\mathrm{R} \times \mathrm{R}_+$. By (Rockafellar, 1970, Theorem 12.2), $\phi^\pi(s,t)$ can therefore be expressed as the conjugate of its conjugate. Note that the conjugate of $\phi^\pi(s,t)$ satisfies

$$\begin{aligned}(\phi^\pi)^*(f,g) &= \sup_{s \in \mathrm{R},\, t \in \mathrm{R}_+} fs + gt - t\phi(s/t) \\ &= \sup_{t \in \mathrm{R}_+} gt + t\phi^*(f) = \begin{cases} 0 & \text{if } f \in \mathrm{dom}(\phi^*) \text{ and } g + \phi^*(f) \leq 0, \\ +\infty & \text{otherwise,} \end{cases}\end{aligned}$$

for all $f, g \in \mathrm{R}$. The second equality in the above expression follows from (**?**, Theorem 16.1). As $\phi^\pi(s,t) = \sup_{f,g \in \mathrm{R}} sf + tg - (\phi^\pi)^*(f,g)$ by virtue of (**?**, Theorem 12.2), the ϕ-divergence is thus given by

$$\begin{aligned}D_\phi(\mu, \hat{\mu}) &= \int_{\mathrm{X}} \sup_{f,g \in \mathrm{R}} \left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\rho}(z) \cdot f + \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z) \cdot g - (\phi^\pi)^*(f,g) \right\} \mathrm{d}\rho(z) \\ &= \int_{\mathrm{X}} \sup_{f \in \mathrm{dom}(\phi^*)} \left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\rho}(z) \cdot f - \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z) \cdot \phi^*(f) \right\} \mathrm{d}\rho(z) \\ &= \sup_{f \in \mathcal{F}} \int_{\mathrm{X}} \left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\rho}(z) \cdot f(z) - \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z) \cdot \phi^*(f(z)) \right\} \mathrm{d}\rho(z),\end{aligned}$$

where the second equality exploits our explicit formula for $(\phi^\pi)^*$ derived above, while the third equality follows from (Rockafellar and Wets, 2009, Theorem 14.60). This theorem applies because the function $h : \mathrm{dom}(\phi^*) \times \mathrm{X} \to \mathrm{R}$ defined through

$$h(f,z) = \frac{\mathrm{d}\mu}{\mathrm{d}\rho}(z) \cdot f - \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\rho}(z) \cdot \phi^*(f)$$

is continuous in f and Borel measurable in z, thus constituting a Carathéodory integrand in the sense of (Rockafellar and Wets, 2009, Example 14.29). The claim then follows immediately from the definition of Radon-Nikodym derivatives. □

CHAPTER 5 Statistical Learning & Robustness

Statistical Learning deals with estimation in the broad sense (parameters of a distribution, best fitting model, etc) from random samples of the source distribution. In order to translate the sample performance to the true performance (on the source distribution), a central question is to evaluate the discrepancy between the sample and the source distributions. In this chapter, we provide such concentration problems for different metrics and connect them to robust optimization on measures.

In the whole chapter, we suppose that $X_1, \dots, X_n \sim_{\text{iid}} \mu$, where μ is a probability measure on the unit hypercube $[0,1]^d$.

The *empirical measure* is defined to be the measure

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

and thus its self random.

In statistics, estimators are constructed as functions of the random sample $(X_1, \dots, X_n)$ (or equivalently as linear functionals of μ_n) and their performance is compared with what could be achieved if μ was known.

Thus, we are in front of a problem of minimization of the average error where the distribution of interest is not perfectly known, i.e., a distributionally robust optimization problem. To control the uncertainty on the distribution, one needs to control the convergence of μ_n to μ. This is the topic of the present chapter.

5.1 Statistical Learning & Convergence of distributions

5.1.1 The laws of large numbers

The strong law of large numbers state that $\frac{1}{n}\sum_{i=1}^{n} X_i = \langle I, \mu_n \rangle \to \mathbb{E}X = \langle I, \mu \rangle$ almost surely (where I is the identity).

If we want to have an idea on how fast this convergence occurs, we can use the fact that in dimension $d = 1$

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \right\|^2 \leq \frac{\sigma^2}{n}$$

since the X_i are independent of finite variance σ^2 (as they are bounded). By Markov's inequality, this leads to bounds known as Hoeffding's inequality for bounded random variables:

$$\mathbb{P}[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\right\|^2 > t] \leq 2\exp(-2nt)$$

In higher dimension, the same type of bound holds with a variance decay at rate n^{-1} (even in infinite dimensional Hilbert spaces).

5.1.2 True and empirical risk

Let f be a continuous bounded function on $[0,1]^d$, for instance standing for some estimation error function.

The true risk is defined as $R(f) = \mathbb{E}_{X\sim\mu}[f(X)] = \langle f, \mu\rangle$ and the empirical risk as $\hat{R}_n(f) = \mathbb{E}_{X\sim\mu_n}[f(X)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \langle f, \mu_n\rangle$.

We can use exactly the same techniques as above to obtain concentration bound of the empirical risk towards the true risk depending on the properties of f.

When the estimator, i.e., the function, f is fixed as is the case in standard parametric statistics this is usually enough. However, in statistical learning, the predictor is often learnt from within a class of functions, it is necessary to decouple the estimation from the error.

Let $\mathcal{F}$ be a class of functions from $[0,1]^d$ to $\mathbb{R}$, we let[20]

[20] We assume their existence. A good exercise is to find conditions on $\mathcal{F}$ for which they exist.

$$f^\star = \operatorname{argmin}_{f\in\mathcal{F}} R(f) \text{ (the best predictor achievable from the source distribution)}$$
$$\hat{f}_n = \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_n(f) \text{ (the best predictor achievable from empirical data)}$$

where we notice that $f^\star$ is deterministic while $\hat{f}_n$ is random (as a function of the random sample) then the *generalization error* is

$$\begin{aligned}
0 &\leq \mathbb{E}_n[R(\hat{f}_n)] - R(f^\star) \\
&= \mathbb{E}_n[R(\hat{f}_n) - \hat{R}_n(f^\star)] \\
&\quad + \mathbb{E}_n[\hat{R}_n(\hat{f}_n) - \hat{R}_n(f^\star)] \ (\leq 0) \\
&\quad + \mathbb{E}_n[\hat{R}_n(f^\star)] - R(f^\star) \\
&\leq 2\mathbb{E}_n \sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|
\end{aligned}$$

where $\mathbb{E}_n$ denotes the expectation over $(X_1, .., X_n)$.

Hence, an important task in statistical learning is to control

$$\mathbb{E}_n \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right|$$

which is done though various techniques: Rademacher complexity, VC-dimension (in classification), metric entropy. We will focus here on the latter. It can also be seen as the average integral probability metric between μ and μ_n.

5.1.3 Classes of functions & distances between distributions

The random process $f \mapsto \frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\}$ is known as an *empirical process* and a standard technique to handle it is though *chaining*. Given a class $\mathcal{F}$ of real-valued functions on $\Omega \subseteq \mathrm{R}^d$, we call a set $F = \{f_1, \dots, f_N\}$ an ε-cover of $\mathcal{F}$ if, for any $f \in \mathcal{F}$, there exists $f_i \in F$ such that $\|f - f_i\|_{L^\infty(\Omega)} \leq \varepsilon$. The *$\varepsilon$-covering number* of $\mathcal{F}$ is

$$N(\varepsilon, \mathcal{F}) = \min\{|F| : F \text{ is an } \varepsilon\text{-cover of } \mathcal{F}\}\,.$$

The chaining argument shows that the covering number of a class $\mathcal{F}$ controls the supremum of an empirical process indexed by that set. We use the following version:

Proposition 5.1 ((Van Handel, 2014, Theorem 5.31)). *If $\mathcal{F}$ is a set of real-valued functions on Ω such that $\|f\|_{L^\infty(\Omega)} \leq R$ for all $f \in \mathcal{F}$, then*

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\} \lesssim \inf_{\tau>0}\left\{\tau + \frac{1}{\sqrt{n}}\int_\tau^R \sqrt{\log N(\varepsilon, \mathcal{F})}\,\mathrm{d}\varepsilon\right\}\,.$$

The above bound is usually not relevant whenever Ω is not compact as $N(\varepsilon, \mathcal{F}) = +\infty$. However, when Ω is a compact set, this bound can lead to performing guarantees.

Example 5.2 (Linear models). As an exercise. Quick convergence but not a metric between distributions.

5.1.4 Conclusion

Statistics rely on non-parametric distances between empirical and source distributions. Among a several candidates, KL is impossible, IPMs seem privileged, W_1 is a good compromise.

5.2 Statistical Optimal Transport

Warning: this section is currently directly an excerpt of (Chewi et al., 2024, Chap. 2) containing the results seen in class.

Since $\int \|\cdot\|\,\mathrm{d}\mu_n \to \int \|\cdot\|\,\mathrm{d}\mu$ almost surely, we have that $W_1(\mu_n, \mu) \to 0$. Moreover, since $W_1(\mu_n, \mu)$ is bounded almost surely, we also have convergence in mean:

$$\mathbb{E}W_1(\mu_n, \mu) \to 0\,.$$

We focus in this part on the rate of this convergence and show that

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{\mathrm{d}} \cdot \begin{cases} n^{-1/2} & \text{if } \mathrm{d} = 1, \\ (\log n/n)^{1/2} & \text{if } \mathrm{d} = 2, \\ n^{-1/\mathrm{d}} & \text{if } \mathrm{d} \geq 3, \end{cases}$$

and that this rate is unimprovable in general.

We observe that the convergence of μ_n to μ in Wasserstein distance degrades exponentially as the dimension grows, a phenomenon often known as the curse of dimensionality. This is due to the fact that W_1 captures the weak convergence of measures and not only the convergence of the generalization error.

5.2.1 Upper bound

First, we show the following upper bound:

Proposition 5.3. *If the support of μ lies in $[0,1]^d$, then*

$$\mathbb{E}W_1(\mu_n,\mu) \lesssim \sqrt{\mathrm{d}} \cdot \begin{cases} n^{-1/2} & \text{if } \mathrm{d}=1, \\ (\log n)\, n^{-1/2} & \text{if } \mathrm{d}=2, \\ n^{-1/\mathrm{d}} & \text{if } \mathrm{d}\geq 3. \end{cases}$$

Using the dual representation of the 1-Wasserstein distance, we can write

$$\begin{aligned} W_1(\mu_n,\mu) &= \sup_{f\in \mathrm{Lip}_1} \left\{ \int f \,\mathrm{d}\mu_n - \int f \,\mathrm{d}\mu \right\} \\ &= \sup_{f\in \mathrm{Lip}_1} \frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\}\,. \end{aligned} \tag{5.1}$$

Proposition 5.1 and (5.1) imply that we can obtain an upper bound on $\mathbb{E}W_1(\mu_n,\mu)$ as long as we can calculate the covering numbers of the set of Lipschitz functions on $\Omega = [0,1]^d$. We also notice that we can assume without loss of generality that the functions appearing in (5.1) take the value 0 at $(0,\dots,0)$. Indeed, a Lipschitz function on $[0,1]^d$ is bounded, and since the value of $\frac{1}{n}\sum_{i=1}^n\{f(X_i)-\mathbb{E}f(X_i)\}$ is unaffected if we shift f by a constant, we may fix its value at $(0,\dots,0)$ to be 0 without loss of generality.

Lemma 5.4. *Denote by $\mathrm{Lip}_1([0,1]^d)$ the set of 1-Lipschitz functions on $[0,1]^d$ satisfying $f(0)=0$. Then*

$$\log N(\varepsilon, \mathrm{Lip}_1([0,1]^d)) \lesssim (4\sqrt{\mathrm{d}}/\varepsilon)^d\,.$$

Proof. We bound the covering number by exhibiting an ε-cover of $\mathrm{Lip}_1([0,1]^d)$ of the specified size. To do so, we again use the notion of a dyadic partition of $[0,1]^d$ into a set Q_j of cubes of side length 2^{-j}. Each element of Q_j is of the form $2^{-j}([k_1,k_1+1]\times\ldots\times[k_\mathrm{d},k_\mathrm{d}+1])$ for some integers $k_1,\dots,k_\mathrm{d}\in[2^j-1]:=\{0,\dots,2^j-1\}$, and we denote such an element by $Q_{\vec{k}}$ for $\vec{k}=(k_1,\dots,k_\mathrm{d})$.[21]
Fix an integer $j\geq 0$ and positive $\delta>0$ to be specified. Consider the set $\mathcal{H}$ of functions h satisfying the following requirements:

1. h is constant on each element of Q_j, i.e., there exist constants $(h_{\vec{k}})_{\vec{k}\in[2^j-1]^d}$ such that $h(x)=h_{\vec{k}}$ for all $x\in Q_{\vec{k}}$.
2. $h_{\vec{k}}$ is an integer multiple of δ for all $\vec{k}\in[2^j-1]^d$.
3. $h_{(0,\dots,0)}=0$.
4. If $\|\vec{k}-\vec{k}'\|_\infty\leq 1$, then $|h_{\vec{k}}-h_{\vec{k}'}|\leq 2^{-j}\sqrt{\mathrm{d}}+\delta$.

We first claim that $\mathcal{H}$ constitutes an ε-cover of $\mathrm{Lip}_1([0,1]^d)$ if $2^{-j}\sqrt{\mathrm{d}}+\delta\leq\varepsilon$. Given any $f\in\mathrm{Lip}_1([0,1]^d)$, denote by h_f the element of $\mathcal{H}$ given by $(h_f)_{\vec{k}}=\delta\lfloor f(2^{-j}(k_1,\dots,k_\mathrm{d}))/\delta\rfloor$ for all $\vec{k}\in[2^j-1]^d$. To see that $h_f\in\mathcal{H}$, note that it immediately satisfies the first three requirements by construction, and for the fourth,

[21] This collection of cubes overlaps at the boundaries, but as above we may remove overlaps to obtain a disjoint partition of $[0,1]^d$.

we have

$$\begin{aligned}|(h_f)_{\vec{k}} - (h_f)_{\vec{k}'}| &= \delta \left| \lfloor f(2^{-j}(k_1, \ldots, k_{\mathrm{d}}))/\delta \rfloor - \lfloor f(2^{-j}(k_1', \ldots, k_{\mathrm{d}}'))/\delta \rfloor \right| \\ &\le |f(2^{-j}(k_1, \ldots, k_{\mathrm{d}})) - f(2^{-j}(k_1', \ldots, k_{\mathrm{d}}'))| + \delta \\ &\le 2^{-j} \|\vec{k} - \vec{k}'\|_2 + \delta,\end{aligned}$$

where the last inequality follows from the fact that f is Lipschitz. Since $\|\vec{k}-\vec{k}'\|_2 \le \sqrt{\mathrm{d}}$ when $\|\vec{k} - \vec{k}'\|_\infty = 1$, the claim follows. Finally, for any $x \in Q_{\vec{k}}$, the fact that f is Lipschitz again implies

$$\begin{aligned}|f(x) - (h_f)_{\vec{k}}| &= \left| f(x) - \delta \lfloor f(2^{-j}(k_1, \ldots, k_{\mathrm{d}}))/\delta \rfloor \right| \\ &\le |f(x) - f(2^{-j}(k_1, \ldots, k_{\mathrm{d}}))| + \delta \\ &\le \operatorname{diam}(Q_{\vec{k}}) + \delta \\ &= 2^{-j}\sqrt{\mathrm{d}} + \delta.\end{aligned}$$

Therefore $\|f - h_f\|_\infty \le 2^{-j}\sqrt{\mathrm{d}} + \delta$.
We have shown that for every $f \in \mathrm{Lip}_1([0,1]^d)$, there exists $h_f \in \mathcal{H}$ such that $\|f - h_f\|_\infty \le 2^{-j}\sqrt{\mathrm{d}} + \delta$. Therefore, if $2^{-j}\sqrt{\mathrm{d}} + \delta \le \varepsilon$, then $\mathcal{H}$ is an ε-cover of $\mathrm{Lip}_1([0,1]^d)$. We fix $\delta = 2^{-j}\sqrt{\mathrm{d}}$, so that this requirement reduces to $2^{-j}\sqrt{\mathrm{d}} \le \varepsilon/2$.
To bound $|\mathcal{H}|$, note that if we fix the value of $h_{\vec{k}}$ for some $\vec{k}$, then for any $\vec{k}'$ such that $\|\vec{k} - \vec{k}'\|_\infty = 1$, there are at most 5 possible values of $h_{\vec{k}'}$. This follows from the fact that $h_{\vec{k}'}$ must be an integer multiple of $\delta = 2^{-j}\sqrt{\mathrm{d}}$, and there are 5 integer multiples of δ in the interval $[h_{\vec{k}} - 2\delta, h_{\vec{k}} + 2\delta]$. Therefore, if we consider specifying an element $\mathcal{H}$ by specifying the values of $h_{\vec{k}}$ sequentially by setting $h_{(0,\ldots,0)} = 0$ and proceeding in lexicographic order, then at each stage we have at most 5 choices for the next value of $h_{\vec{k}}$. This implies that $|\mathcal{H}| \le 5^{2^{dj}-1}$.
For any j for which $2^{-j}\sqrt{\mathrm{d}} \le \varepsilon/2$, we have therefore obtained an ε-cover $\mathcal{H}$ of $\mathcal{F}$ satisfying $\log|\mathcal{H}| \lesssim 2^{\mathrm{d}j}$. Choosing 2^j to be the smallest power of two larger than $2\sqrt{\mathrm{d}}/\varepsilon$ yields the claim. □

With the bound of Lemma 5.4 in hand, we can give another proof of Proposition 5.3.

Proof of Proposition 5.3. Since $\|f\|_\infty \le \sqrt{\mathrm{d}}$ for all $f \in \mathrm{Lip}_1([0,1]^d)$, by Proposition 5.1 and (5.1), for any $\tau > 0$,

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \tau + \frac{1}{\sqrt{n}} \int_\tau^{\sqrt{\mathrm{d}}} \sqrt{\log N(\varepsilon, \mathrm{Lip}_1([0,1]^d))}\, \mathrm{d}\varepsilon.$$

Applying Lemma 5.4 yields

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \tau + \frac{1}{\sqrt{n}} \int_\tau^{\sqrt{\mathrm{d}}} (4\sqrt{\mathrm{d}}/\varepsilon)^{\mathrm{d}/2}\, \mathrm{d}\varepsilon.$$

We now consider the bound separately for $\mathrm{d} = 1$ and $\mathrm{d} > 1$. If $\mathrm{d} = 1$, then we may take $\tau = 0$ to obtain

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \frac{1}{\sqrt{n}} \int_0^1 (4/\varepsilon)^{1/2}\, \mathrm{d}\varepsilon \lesssim n^{-1/2}.$$

If $\mathrm{d} > 1$, then $\varepsilon^{-\mathrm{d}/2}$ is no longer integrable at 0, so we take $\tau = 4\sqrt{\mathrm{d}}\, n^{-1/\mathrm{d}}$ to obtain

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{\mathrm{d}}\, n^{-1/\mathrm{d}} + \frac{1}{\sqrt{n}} \int_{4\sqrt{\mathrm{d}}\, n^{-1/\mathrm{d}}}^{\sqrt{\mathrm{d}}} (4\sqrt{\mathrm{d}}/\varepsilon)^{\mathrm{d}/2} \, \mathrm{d}\varepsilon \,.$$

When $\mathrm{d} = 2$, the integral is $O(\log n)$, and we obtain $\mathbb{E}W_1(\mu_n, \mu) \lesssim (\log n)/\sqrt{n}$. When $\mathrm{d} > 2$, the integral is $O(n^{1/2-1/\mathrm{d}})$, and we obtain $\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{\mathrm{d}}\, n^{-1/\mathrm{d}}$. □

5.2.2 *Optimality*

We have established upper bounds on the Wasserstein distance between the empirical distribution μ_n and the data generating distribution μ and shown rates of order $n^{-1/\mathrm{d}}$. While this result readily yields consistency, the rate is slow even in moderate dimensions and is symptomatic of the curse of dimensionality that plagues most non-parametric methods. One could wonder then whether such rates can be improved.

While a negative answer to the second question implies a negative answer to the first one—if no estimator can estimate μ faster than $n^{-1/\mathrm{d}}$ then certainly the empirical measure μ_n cannot—we also make the negative answer to the first question explicit since it is, in some sense stronger. Indeed, we show below that even in the case where μ is the uniform measure on $[0,1]^d$ then, $\mathbb{E}[W_1(\mu_n, \mu)] \gtrsim n^{-1/\mathrm{d}}$. However, in that case, there is clearly a better estimator than μ_n: simply take $\tilde{\mu}_n = \mu$ itself! The answer to the second question relies on the theory of minimax lower bounds as in (Tsybakov, 2009, Chapter 2) and states that for any estimator, i.e., any measurable function $\tilde{\mu}_n = \tilde{\mu}_n(X_1, \dots, X_n)$ of the data $X_1, \dots, X_n$, there exists μ supported on $[0,1]^d$ such that $\mathbb{E}[W_1(\tilde{\mu}_n, \mu)] \gtrsim n^{-1/\mathrm{d}}$. Unlike the lower bound for the empirical measure μ_n, in the minimax lower bounds, the unfavorable distribution μ is not explicit.

Lower bounds for the empirical measure μ_n

The goal of this section is to show that any distribution supported on n points has to be far from the uniform measure on $[0,1]^d$ in W_1 distance.

Theorem 5.5. *Fix $\mathrm{d} \geq 3$ and let μ denote the uniform measure on $[0,1]^d$. Then for any measure $\tilde{\mu}_n$ supported on n points $x_1, \dots, x_n \in \mathrm{R}^d$, it holds*

$$W_1(\tilde{\mu}_n, \mu) \geq \frac{1}{108\mathrm{d}}\, n^{-1/\mathrm{d}} \,.$$

Proof. We employ the Kantorovich-Rubinstein formulation so that proving a lower bound on W_1 can be done by exhibiting a 1-Lipschitz function with the desired property. Given $x \in [0,1]^d$, let $\xi(x) \in \{x_1, \dots, x_n\}$ denote the closest point to x in $\{x_1, \dots, x_n\}$ (ties are broken arbitrarily). Next, consider the function

$$f_n(x) = \|x - \xi_n(x)\| \,,$$

which is 1-Lipschitz thanks to the reverse triangle inequality. Moreover, for any $i = 1, \dots, n$, we have $f_n(x_i) = 0$ so that $\int f \, \mathrm{d}\tilde{\mu}_n = 0$. Hence

$$W_1(\tilde{\mu}_n, \mu) \geq \int f_n \, \mathrm{d}\mu = \int \|x - \xi_n(x)\| \, \mu(\mathrm{d}x) \,.$$

To bound this quantity from below, we show that μ places significant mass on points that are far from *any* x_i. To that end, consider a partition Q of $[0,1]^d$ into cubes of

side length $(2n)^{-1/\mathrm{d}}$. Since $|\mathcal{Q}| = 2n$, there exist n such cubes $Q_1, \dots, Q_n$ that do not contain any of the x_i's. Let $Q \in \mathcal{Q}$ be one such cube with center q and consider its subcube $Q' \subset Q$ also with center q but with a smaller side length than Q by a factor of $1 - 2/\mathrm{d}$. Using Minkowski sum notation, we can write this as:

$$Q' = \left(1 - \frac{2}{\mathrm{d}}\right)(Q - \{q\}) + \{q\}\,.$$

By construction, any $x \in Q'$ satisfies

$$\|x - \xi_n(x)\| \geq \inf_{\substack{x \in Q' \\ y \in Q^c}} \|x - y\| = \frac{1}{\mathrm{d}} \cdot (2n)^{-1/\mathrm{d}}\,.$$

Hence

$$\int \|x - \xi_n(x)\|\, \mu(\mathrm{d}x) \geq \sum_{i=1}^{n} \int_{Q_i'} \|x - \xi_n(x)\|\, \mu(\mathrm{d}x) \geq \frac{(2n)^{-1/\mathrm{d}}}{\mathrm{d}} \sum_{i=1}^{n} \mu(Q_i')\,.$$

We conclude by observing that

$$\mu(Q_i') = \left(\frac{1 - 2/\mathrm{d}}{(2n)^{1/\mathrm{d}}}\right)^d \geq \frac{1}{54n}\,,$$

where we used the fact that $\mathrm{d} \mapsto (1 - 2/\mathrm{d})^d$ is increasing and that $\mathrm{d} \geq 3$. □

Theorem 5.5 shows that $W_1(\mu_n, \mu)$ is indeed of order $n^{-1/\mathrm{d}}$ at least for $d \geq 3$. In fact the lower bound holds almost surely in $X_1, \dots, X_n$ since it only exploits the fact that μ_n has a support of size at most n.

Minimax lower bounds

While it is hard to think of a better estimator for μ than μ_n in general it could be the case that there exists another estimator $\tilde{\mu}_n$ for which $\mathbb{E}[W_1(\tilde{\mu}_n, \mu)]$ is smaller than $\mathbb{E}[W_1(\mu_n, \mu)]$ uniformly over all measures μ. This possibility is ruled out by the following minimax lower bound.

Theorem 5.6. *Fix $\mathrm{d} \geq 3, n \geq 8$ and let $X_1, \dots, X_n$ be n i.i.d. observations from a distribution μ on R^d. For any estimator $\tilde{\mu}_n$, i.e., any measurable function of $X_1, \dots, X_n$, there exists a measure μ supported on $[0,1]^d$ such that*

$$\mathbb{E}_\mu[W_1(\tilde{\mu}_n, \mu)] \geq \frac{1}{16}\,(2n)^{-1/\mathrm{d}}\,.$$

Proof. Our proof relies on classical techniques for minimax lower bounds. In particular, we use Theorem 2.12 in (Tsybakov, 2009). According to this theorem, if we can find 2^m probability measures indexed by $\omega \in \{-1,1\}^m$ each supported on $[0,1]^d$ such that

(i) $W_1(\mu^{(\omega)}, \mu^{(\omega')}) \geq \frac{r_n}{2} \sum_{j=1}^{m} |\omega_j - \omega_j'|$ for any $\omega, \omega' \in \{-1,1\}^m$,

(ii) for any $\omega, \in \{-1,1\}^m$ differing in at most one coordinate,

$$\mathsf{KL}(\mu^{(\omega)} || \mu^{(\omega')}) \leq \frac{1}{2n}\,,$$

then for any estimator $\tilde{\mu}_n$ based on n i.i.d. observations, there exists $\omega \in \{-1,1\}^m$

such that

$$\mathbb{E}_{\mu^{(\omega)}}[W_1(\tilde{\mu}_n, \mu^{(\omega)})] \geq \frac{mr_n}{4}.$$

In our construction, we take $m = n$ and define the measures $\mu^{(\omega)}$ to be supported on a discrete set as follows. As in the proof of Theorem 5.5, let Q denote a partition of $[0,1]^d$ into $2n$ cubes of side length $(2n)^{-1/\mathrm{d}}$ and let $q_1, \ldots, q_{2n}$ denote their centers. Let $\mu^{(0)}$ denote the uniform measure on $\{q_1, \ldots, q_{2n}\}$:

$$\mu^{(0)} = \frac{1}{2n} \sum_{i=1}^{2n} \delta_{q_i}.$$

For $\omega \in \{-1,1\}^n$, let $\mu^{(\omega)}$ denote a perturbation of $\mu^{(0)}$ defined as

$$\mu^{(\omega)} = \mu^{(0)} + \frac{\alpha}{2n} \sum_{i=1}^{n} \omega_i \left(\delta_{q_i} - \delta_{q_{n+i}}\right),$$

where $\omega = (\omega_1, \ldots, \omega_n)$ and $\alpha \in (0,1)$ is to be defined later. Note that $\mu^{(\omega)}$ is a probability measure.
Since $\|q_j - q_k\| \geq (2n)^{-1/\mathrm{d}}$ for $j \neq k$ for we have

$$W_1(\mu^{(\omega)}, \mu^{(\omega')}) \geq \frac{\alpha}{2n} (2n)^{-1/\mathrm{d}} \sum_{j=1}^{n} |\omega_j - \omega'_j| =: \frac{r_n}{2} \sum_{j=1}^{n} |\omega_j - \omega'_j|$$

for any $\omega, \omega' \in \{0\}^n \cup \{-1,1\}^n$.
It remains to show that (ii) holds for a suitable choice of α. To that end, suppose that ω and ω' differ on the jth coordinate. Observe that

$$\begin{aligned}
\mathrm{KL}(\mu^{(\omega)} || \mu^{(\omega')}) &= \sum_{i=1}^{2n} \mu^{(\omega)}(q_i) \log\left(\frac{\mu^{(\omega)}(q_i)}{\mu^{(\omega')}(q_i)}\right) \\
&= \frac{1}{2n} \left\{ (1+\alpha\omega_j) \log \frac{1+\alpha\omega_j}{1-\alpha\omega_j} + (1-\alpha\omega_j) \log \frac{1-\alpha\omega_j}{1+\alpha\omega_j} \right\} \\
&= \frac{\alpha}{n} \log \frac{1+\alpha}{1-\alpha},
\end{aligned}$$

and this quantity is smaller than $\frac{1}{2n}$ if $\alpha = \frac{1}{4}$. With this choice of α, we obtain

$$r_n = \frac{1}{4n} (2n)^{-1/d},$$

which implies the desired bound. □

5.3 Distributional Robust Optimization

Distributionally Robust Optimization studies decision problems under uncertainty where the probability distribution governing the uncertain problem parameters is itself uncertain, in particular because it is only known though samples. A key component of any DRO model is its ambiguity set, that is, a family of probability distributions consistent with any available structural or statistical information.

This part is mainly based on the monograph of (Kuhn et al., 2024).

5.3.1 Motivation

We will place ourselves in a classical machine learning setting where we have access to n data points $(X_i)_{i=1}^n$ that can stand for observed situations (e.g. past stock prices in portfolio selection, electricity consumptions and weather conditions in energy planning) or labeled training data of the form $X_i = (x_i, y_i)$ (in classical classification and regression problems).

We are interested here in providing a statistical learning method that is performing and reliable on future, unseen situations. To do so, we have to cleverly select a model among a family parametrized by $x \in \mathcal{X}$. We suppose that we have access to the loss $f(x; X)$ (real-valued, lower is better) suffered by the model parametrized by x when facing the situation $X \in \mathrm{X}$. By this, we mean that for all parameters x and situations X we have an explicit and implementable[22] expression for $f(x; X)$.

[22] In the project, we will need to compute derivatives of $f(x; X)$ with respect to x and X numerically, preferably by automatic differentiation.

An ideal way to choose a model would be to select the one with smallest (expected) error in the target application. Yet, while we do know the distribution of the samples $\mu_\mathrm{n} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, we have no access to (the distribution of) future situations. This is where *robust optimization* and *statistical learning* come into play.

Using the data, the Empirical Risk Minimization (aka Sample Average Approximation (Shapiro et al., 2021, Chap. 5)) seeks to minimize the *expected loss over the empirical distribution* i.e.,

$$\min_{x\in\mathcal{X}} \mathbb{E}_{X\sim\mu_\mathrm{n}}[f(x;X)] = \frac{1}{n}\sum_{i=1}^n f(x;X_i)\,. \tag{5.2}$$

ERM is generally regarded as the standard baseline for model training in machine learning. Nonetheless, the approach is built over the assumption that the empirical distribution μ_n is close to the distribution met in the target application. While this is verified in some applications, the goal of our project is to focus on cases where this is not necessarily the case. For instance, we may have too few samples to approximate correctly their underlying distribution, or we may face a distribution shift between training and application of the model.

Also, ERM fails to provide a good estimation of the future performance of a model with the training error. Indeed, if the samples are drawn independently from the same distribution μ_true, classical statistical learning theory ensures that, with high probability, $\mathbb{E}_{X\sim\mu_\mathrm{true}}[f(x;X)]$ is *close to* $\mathbb{E}_{X\sim\mu_\mathrm{n}}[f(x;X)]$ up to $O\big(1/\sqrt{n}\big)$ error terms; see Proposition 5.3, (Boucheron et al.; Wainwright, 2019), this kind of relation is known as a *generalization* result. But being *close* does not *guarantee* any performance, even facing the underlying distribution μ_true. We have not much control over the probability that $\mathbb{E}_{X\sim\mu_\mathrm{n}}[f(x;X)] < \mathbb{E}_{X\sim\mu_\mathrm{true}}[f(x;X)]$, i.e., that the real loss is higher than the training loss.

Distributionally Robust Optimization (DRO) Between the pessimistic worst case approach and the classical ERM, a middle spot has to be found. To do so, we can reasonably acknowledge that the empirical distribution provides *partial* information about the encountered distribution of X in practice, i.e., that the two distributions are close. Doing so, *we depart from pointwise robustness to consider distributional robustness.* DRO thus consists in minimizing the *worst expectation of the loss* when the *distribution* lives in a neighborhood $\mathcal{U}(\mu_\mathrm{n})$ of μ_n. The resulting problem is

$$\min_{x\in\mathcal{X}} \sup_{\nu\in\mathcal{U}(\mu_\mathrm{n})} \mathbb{E}_{X\sim\nu}[f(x;X)] \tag{5.3}$$

where the inner sup is taken over *probability measures* on X in the set $\mathcal{U}(\mu_n)$. First, we can notice that if $\mathcal{U}(\mu_n)$ is reduced to the singleton $\{\mu_n\}$, the problem is equivalent to (5.2). More interestingly, if $\mu_{\text{true}} \in \mathcal{U}(\mu_n)$, then the optimal value of Problem (5.3) is an *upper-bound on* $\mathbb{E}_{X\sim\mu_{true}}[f(x;X)]$, i.e., an *exact generalization bound* that precisely match our quest for predictability in the performance of machine learning models. However, if the ambiguity set is too loose, the distributions can become unfavorable (maybe including discrete ones centered on worst case points), and we fall back to the caveat of (pointwise) worst-case robustness, hence the difficulty to design $\mathcal{U}(\mu_n)$. In addition, since the inner maximization over probability measures is an *infinite-dimensional problem*, a *compromise* has to be found between the *modelling capacity* and the *computational tractability* of the objective.

Wasserstein Distributionally Robust Optimization In order to take into account situations that are outside of the already observed ones and in order to encompass both absolutely continuous and discrete distributions in a common neighborhood, a natural approach is to rely on the Wasserstein distance, originating from optimal transport (Villani et al., 2009, Chap. 6). This approach leads to ambiguity sets of the form $\mathcal{U}(\mu_n) = \{\nu \in \mathcal{P}(\mathrm{X}) : W(\mu_n, \nu) \leq \rho\}$ for some $\rho \geq 0$, where $\mathcal{P}(\mathrm{X})$ is the set of probability distributions on X and, for a lower semi-continous cost function $c : \mathrm{X}\times\mathrm{X} \to \mathrm{R}_+$, the Wasserstein distance[23] between μ_n and ν is defined as the optimal transport cost between the two measures:

$$W_c(\mu_n, \nu) = \inf\left\{\mathbb{E}_{(X,Y)\sim\pi}[c(X,Y)] : \pi \in \mathcal{P}(\mathrm{X}\times\mathrm{X}), \pi_1 = \mu_n, \pi_2 = \nu\right\},$$

with π_1 (resp. π_2) the first (resp. second) marginal of the transport plan π. With such an ambiguity set, the DRO Problem (5.3) becomes a *Wasserstein Distributionally Robust (WDRO) problem*. We refer to (Blanchet et al., 2023) for a recent review of WDRO and connections with DRO.

[23] In order to match the literature's terminology, we will abusively call the optimal transport cost the Wasserstein distance even though it is not necessarily a distance when c is not distance-based.

An important point here is that the transport cost plays a crucial role in uncertainty modelling. Classical costs in optimal transport include the norm of the difference $c(X,Y) = \|X - Y\|$ leading to the type-1 Wasserstein distance and the *squared* norm of the difference $c(X,Y) = \|X - Y\|^2$ leading to the type-2 Wasserstein distance *squared* (Villani et al., 2009, Chap. 6). These two choices lead to actual distances in the spaces of measures, which will have an important role for studying statistical properties and distribution shifts. Though natural, these choices are not always suited for the situations encountered in machine learning. Typically, in binary classification tasks, data points are of the form $X = (x, y) \in \mathbb{R}^d \times \{0, 1\}$ and thus the uncertainty in x is very different from the one in y. For such cases, it is useful to define transport costs of the form $c(X = (x, y), Y = (x', y')) = \|x - x'\| + \kappa\mathbb{1}_{y\neq y'}$ with $\kappa > 0$ and $\mathbb{1}_{y\neq y'} = 1$ if $y \neq y'$ and 0 otherwise. Thus, the choice of transport cost is an important aspect to keep in mind when modeling uncertainties with WDRO, which will re-appear in the optimization of the objective.

5.3.2 WDRO problems

We will place ourselves in the following context:

- The objective f is a bounded continuous function (we drop the dependency in x for now)
- We seek robustness around an empirical distribution μ_n consisting of n i.i.d. samples from some distribution μ

- The distributions live in a compact $X \subset \mathbb{R}^d$ (thus their mean, variance is finite) and we rely on the type-1 Wasserstein distance

$$W_1(\mu, \nu) := \inf \left\{ \mathbb{E}_{(X,Y)\sim\pi}[\|X - Y\|] : \pi \in \mathcal{P}(X \times X), \pi_1 = \mu_n, \pi_2 = \nu \right\}$$
$$= \sup_{f \in \mathrm{Lip}_1} \int f \mathrm{d}\mu - \int f \mathrm{d}\nu$$

  where the equality comes from the Kantorovich-Rubinstein duality (see (Villani et al., 2009, Rem. 6.5)).

In this part, we study the problem

$$\sup_{\nu \in \mathcal{U}(\mu_n)} \mathbb{E}_{X\sim\nu}[f(X)] \tag{5.4}$$

where $\mathcal{U}(\mu_n) = \{\mu \in \mathcal{P}(X) : W_1(\mu, \mu_n) \leq \rho\}$

Optimality

The first thing to do is to ensure that this problem has a finite value at that its optimal value is attained.

Theorem 5.7. *The problem (5.4) is finite and its optimum is attained by some probability measure $\mu^\star$.*

Proof. In our setting, since W_1 metrizes the weak convergence of measures (Villani et al., 2009, Th. 6.9), we have that the objective is weakly continuous and the constraint set is weakly closed and thus sequentially compact by Prokhorov's theorem (note that X is compact). Hence, by Weierstrass' theorem, the optimum of the problem is finite and its optimal value is attained. □

There are absolutely continuous distributions in Wasserstein balls and also discrete (atomic distributions), see the proof of (Villani et al., 2009, Th. 6.18).

Nevertheless, in our setting, the optimal solution has at most $n + 1$ atoms.

Theorem 5.8. *The optimal value $\mu^\star$ of (5.4) is concentrated on at most $n + 1$ atoms.*

Proof. See (Pinelis, 2016). □

Regularization effect

Theorem 5.9. *Let f be L-Lipchitz continuous. Then, we have*

$$\sup_{\nu \in \mathcal{U}(\mu_n)} \mathbb{E}_{X\sim\nu}[f(X)] \leq \mathbb{E}_{X\sim\mu_n} f(X) + \rho L$$

Proof. See Theorem 8.5 in (Kuhn et al., 2024). □

Duality

Theorem 5.10. *We have*

$$\sup_{\nu \in \mathcal{U}(\mu_n)} \mathbb{E}_{X\sim\nu}[f(X)] = \inf_{\lambda > 0} \lambda\rho + \mathbb{E}_{X\sim\mu_n} \sup_{Y \in X} f(Y) - \lambda\|X - Y\|$$

Proof. See in class. □

Regularization of WDRO

Exercises

Exercise 5.1 (Kernel embeddings, MMD, and IPM)**.**
Given a symmetric, positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the Moore-Aronszajn theorem asserts the existence of a unique RKHS $\mathcal{H}$ on $\mathcal{X}$ (a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a norm $\| \cdot \|_{\mathcal{H}}$) for which k is a reproducing kernel, i.e., in which the element $k(x, \cdot)$ satisfies the reproducing property

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \qquad \forall f \in \mathcal{H}, \quad \forall x \in \mathcal{X}$$

and in particular, taking $f = k(y, \cdot)$,

$$\langle k(y, \cdot), k(x, \cdot) \rangle_{\mathcal{H}} = k(x, y) \text{ and } \|k(x, \cdot)\|_{\mathcal{H}} = k(x, x)$$

which are both computable quantities (using only k).
One may alternatively consider $x \mapsto k(x, \cdot)$ as an implicit feature mapping $\varphi : \mathcal{X} \to \mathcal{H}$ (which is therefore also called the feature space), so that $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ can be viewed as a measure of similarity between points $x, x' \in \mathcal{X}$. While the similarity measure is linear in the feature space, it may be highly nonlinear in the original space depending on the choice of kernel.
The kernel embedding of the distribution μ in $\mathcal{H}$ (also called the kernel mean or mean map) is given by:

$$m_\mu := \mathbb{E}_{X \sim \mu}[k(X, \cdot)] = \int k(x, \cdot) \mathrm{d}\mu(x) = \mathbb{E}[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) \, \mathrm{d}\mu(x)$$

If μ allows a square integrable density p, then $m_\mu = \mathcal{E}_k p$, where $\mathcal{E}_k$ is the Hilbert-Schmidt integral operator. A kernel is characteristic if the mean embedding $m : \{\text{family of distributions over } \mathcal{X}\} \to \mathcal{H}$ is injective. Each distribution can then be uniquely represented in the RKHS and all statistical features of distributions are preserved by the kernel embedding if a characteristic kernel is used. Finally, note that computationally speaking, we only have access to k and never to elements of $\mathcal{H}$ or to the feature map φ.
The maximum mean discrepancy (MMD) is then defined as

$$MMD(\mu, \nu) = \|m_\mu - m_\nu\|_{\mathcal{H}}.$$

1. Show that MMD is an IPM for $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$.
2. Show that the linear kernel $k(x, y) = \langle x, y \rangle$ is not characteristic.
3. Show that $MMD^2(\mu, \nu) = \iint k(x, y) \mathrm{d}\mu(x) \mathrm{d}\mu(y) + \iint k(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y) - 2 \iint k(x, y) \mathrm{d}\mu(x) \mathrm{d}\nu(y)$

Elements of Solution: For part 1.,

$$\begin{aligned} MMD(\mu,\nu) &= \| \int k(x,\cdot)\mathrm{d}\mu(x) - \int k(x,\cdot)\mathrm{d}\nu(x)\|_{\mathcal{H}} \\ &= \sup_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq 1} \langle \int k(x,\cdot)\mathrm{d}\mu(x) - \int k(x,\cdot)\mathrm{d}\nu(x), f\rangle_{\mathcal{H}} \\ &= \sup_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq 1} \int \langle k(x,\cdot)\,, f\rangle_{\mathcal{H}}\mathrm{d}\mu(x) - \int \langle k(x,\cdot)\,, f\rangle_{\mathcal{H}}\mathrm{d}\nu(x) \\ &= \sup_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq 1} \int f(x)\mathrm{d}\mu(x) - \int f(x)\mathrm{d}\nu(x) \end{aligned}$$

and the absolute value can be added since if $f \in \mathcal{H}, -f \in \mathcal{H}$.
For part 3., The kernel mean embedding of a probability measure μ into $\mathcal{H}$ is defined as

$$m_\mu = \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x).$$

Then,

$$\mathrm{MMD}(\mu,\nu) = \|m_\mu - m_\nu\|_{\mathcal{H}}.$$

Expanding the squared norm gives:

$$\|m_\mu - m_\nu\|_{\mathcal{H}}^2 = \langle m_\mu, m_\mu\rangle_{\mathcal{H}} + \langle m_\nu, m_\nu\rangle_{\mathcal{H}} - 2\langle m_\mu, m_\nu\rangle_{\mathcal{H}}.$$

Using the reproducing property,

$$\langle m_\mu, m_\mu\rangle_{\mathcal{H}} = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\, d\mu(x)\, d\mu(x'),$$

$$\langle m_\nu, m_\nu\rangle_{\mathcal{H}} = \int_{\mathcal{X}}\int_{\mathcal{X}} k(y,y')\, d\nu(y)\, d\nu(y'),$$

and

$$\langle m_\mu, m_\nu\rangle_{\mathcal{H}} = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mu(x)\, d\nu(y).$$

Thus,

$$\mathrm{MMD}^2(\mu,\nu) = \mathbb{E}_{X,X'\sim\mu}[k(X,X')] + \mathbb{E}_{Y,Y'\sim\nu}[k(Y,Y')] - 2\,\mathbb{E}_{X\sim\mu,\,Y\sim\nu}[k(X,Y)].$$

□

Exercise 5.2 (Comparison of Integral Probability Metrics)**.** Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a positive definite kernel with associated Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ and assume that $\sup_{x\in\mathcal{X}} \|\nabla_x k(\cdot,x)\|_{\mathcal{H}} \leq 1$.
Prove that for any pair of probability distributions on $\mathcal{X}$, we have

$$\mathrm{MMD}(\mu,\nu) \leq W_1(\mu,\nu).$$

Elements of Solution: Let $f \in \mathcal{H}$ where $\mathcal{H}$ is the RKHS corresponding to the kernel k. By the reproducing property,

$$f(x) = \langle f, k(\cdot,x)\rangle_{\mathcal{H}}.$$

Differentiating with respect to x yields

$$\nabla f(x) = \langle f, \nabla_x k(\cdot, x) \rangle_{\mathcal{H}}.$$

By the Cauchy–Schwarz inequality,

$$\|\nabla f(x)\| \leq \|f\|_{\mathcal{H}} \, \|\nabla_x k(\cdot, x)\|_{\mathcal{H}}.$$

Thus, for any $f \in \mathcal{H}$,

$$\|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|} \leq \|f\|_{\mathcal{H}}.$$

In particular, if $\|f\|_{\mathcal{H}} \leq 1$, then f is Lipschitz with constant 1 and thus

$$\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\} \subset \{f : \|f\|_{\mathrm{Lip}} \leq 1\}.$$

□

Exercise 5.3. Let $\mathcal{X}$ be a compact metric space and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, positive definite kernel. The *Reproducing Kernel Hilbert Space (RKHS)* $\mathcal{H}$ associated with k is defined as the completion of the linear span of the functions $k(\cdot, x)$ for $x \in \mathcal{X}$. In many applications, one is interested in whether $\mathcal{H}$ is rich enough to approximate all continuous functions on $\mathcal{X}$ uniformly. When this is the case, we say that the kernel k is *universal*.

Definition 5.11 (Universal Kernel)**.** A continuous kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a compact metric space $\mathcal{X}$ is called *universal* if its RKHS $\mathcal{H}$ is dense in $C(\mathcal{X})$ (the space of continuous functions on $\mathcal{X}$) with respect to the uniform norm.

A kernel k is universal if it satisfies the following conditions:

1. **Continuity:** k is continuous on $\mathcal{X} \times \mathcal{X}$.
2. **Strict Positive Definiteness:** For any distinct points $x_1, \ldots, x_n \in \mathcal{X}$ and nonzero coefficients $c_1, \ldots, c_n$,

$$\sum_{i,j=1}^{n} c_i c_j \, k(x_i, x_j) > 0.$$

   This ensures that the mapping $x \mapsto k(\cdot, x)$ is injective.
3. **Separation of Points and Constants:** The linear span $\mathcal{A} = \mathrm{span}\{k(\cdot, x) : x \in \mathcal{X}\}$ separates points in $\mathcal{X}$ and the constant functions are contained in (or can be approximated arbitrarily well by) $\mathcal{A}$.

For example, when $\mathcal{X} \subset \mathbb{R}^d$ is compact and k is translation invariant (i.e., $k(x, y) = \psi(x - y)$) with the Fourier transform of ψ strictly positive everywhere, then k is universal. The Gaussian RBF kernel is a well-known example.
Show that if the RKHS associated with k is sufficiently rich , one may also relate d_{TV} and MMD and show that

$$d_{TV}(\mu, \nu) \leq \mathrm{MMD}(\mu, \nu).$$

This part contains a series of exercises related to concentration and statistical robustness.

Exercise 5.4 (Concentration for the MMD distance)**.** Let k be a characteristic kernel such that $k(x,x) \leq 1$ for any $x \in \mathbb{R}^d$. Let $X_1, \dots, X_n$ be n i.i.d. observations from a distribution μ on $\mathbb{R}^d$ and define the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} .$$

Then

$$\mathbb{E}[\mathrm{MMD}(\mu_n, \mu)] \leq \frac{1}{\sqrt{n}} .$$

Elements of Solution: It follows from Exercise 5.1 that

$$\begin{aligned}
\mathbb{E}[\mathrm{MMD}^2(\mu_n, \mu)] &= \mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^{n} \{k(X_i, \cdot) - \mathbb{E}k(X_i, \cdot)\} \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n} \mathbb{E}\|k(X_1, \cdot) - \mathbb{E}k(X_1, \cdot)\|_{\mathcal{H}}^2 \\
&= \frac{1}{n} \left(\mathbb{E}\|k(X_1, \cdot)\|_{\mathcal{H}}^2 - \|\mathbb{E}k(X_1, \cdot)\|_{\mathcal{H}}^2 \right) \\
&\leq \frac{1}{n} \mathbb{E}\|k(X_1, \cdot)\|_{\mathcal{H}}^2 .
\end{aligned}$$

Next, observe that

$$\mathbb{E}\|k(X_1, \cdot)\|_{\mathcal{H}}^2 = \mathbb{E}[k(X_1, X_1)] \leq 1 .$$

The claim follows from Jensen's inequality. □

Bibliography

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 22(128):1–59, 2021.

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

Aharon Ben-Tal, Arkadi Nemirovski, and Laurent El Ghaoui. Robust optimization. 2009.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 2006.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Jose Blanchet, Daniel Kuhn, Jiajin Li, and Bahar Taskesen. Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*, 2023.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024. URL https://arxiv.org/abs/2407.18163.

Christian Clason and Tuomo Valkonen. Introduction to nonsmooth analysis and optimization. *arXiv preprint arXiv:2001.00216*, 2020.

Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

F Facchinei and J S Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1993a. Two volumes.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1993b. Two volumes.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization, 2024. URL https://arxiv.org/abs/2411.02549.

Claude Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 251(254):10, 2012.

Boris S Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*, volume 330. Springer Science & Business Media, 2006.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Juan Peypouquet. *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.

Iosif Pinelis. On the extreme points of moments sets. *Mathematical Methods of Operations Research*, 83(3):325–349, 2016.

Olivier Rioul. This is it: A primer on shannon's entropy and information. In *Information Theory: Poincaré Seminar 2018*, pages 49–86. Springer, 2021.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, 1998.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387790527. URL https://books.google.fr/books?id=mwB8rUBsbqoC.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100 (1):295–320, 1928.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *Advances in Neural Information Processing Systems*, 34:18866–18877, 2021.

Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over wasserstein balls. *Mathematical Programming*, 195(1):1107–1122, 2022.

APPENDIX A DIFFERENTIABILITY AND SMOOTHNESS

In the first page of the renowned book "Variational analysis" by R. Tyrrell Rockafellar and Roger J-B Wets (Rockafellar and Wets, 1998), we are told that "it's convenient for many purposes to consider functions f that are allowed to be extended-real-valued, i.e., to take values in $\overline{\mathbb{R}} = [-\infty, +\infty]$ instead of just $\mathbb{R} = (-\infty, +\infty)$", we will thus adopt this convention ourselves.

A fundamental question in variational analysis is the study of the minimum (or equivalently maximum) of functions defined over a Euclidean space $\mathbb{R}^n$. In all this course, we will place ourselves in the (finite-dimensional) Euclidean space $\mathbb{R}^n$, with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $x \mapsto \|x\| := \sqrt{\langle x, x \rangle}$.

For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, we define its *domain* as $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$, and its *infimum*

$$\inf f := \inf_{x \in \mathbb{R}^n} f(x) = \inf_{x \in \operatorname{dom} f} f(x).$$

Whenever this infimum is attained, ie. there is some x such that $f(x) = \inf f$, then it is called a minimum and is denoted by $\min f$. We further define

$$\operatorname{argmin} f := \{x \in \mathbb{R}^n : f(x) = \inf f\}.$$

Additionally, a function f is *lower semi-continuous* if for any $x \in \mathbb{R}^n$,

$$\liminf_{u \to x} f(u) := \min\{t \in \overline{\mathbb{R}} : \exists u_r \to x \text{ with } f(u_r) \to t\} = f(x).$$

Finally, a function f is said to be *proper* is $f(x) < +\infty$ for at least one $x \in \mathbb{R}^n$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. This means that the domain of a proper function is a nonempty set over which f is finite-valued.

A.1 SUBGRADIENTS

In order to investigate the local behavior of a function with respect to minimization, a first natural step is to consider local affine lower approximations. This *first-order* information is captured by the notion of subgradients. There is a variety of subgradients and several ways to express them, see (Rockafellar and Wets, 1998, Chap. 7,8), (Mordukhovich, 2006, Chap. 1) for general references. We give here only the notions that will be used for our purposes following the terminology and notations of (Rockafellar and Wets, 1998, Chap. 8).

Definition A.1 (Subgradients)**.** Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which $f(x)$ is finite:

- the set of *regular subgradients* is defined as

$$\widehat{\partial} f(x) = \{v : f(u) \geq f(x) + \langle v, u - x\rangle + o(\|u - x\|) \text{ for all } u \in \mathbb{R}^n\}\,. \tag{A.1}$$

- the set of *(general or limiting) subgradients* is defined as

$$\partial f(x) = \left\{\lim_r v_r : v_r \in \widehat{\partial} f(u_r),\ u_r \to x,\ f(u_r) \to f(x)\right\}. \tag{A.2}$$

If $f(x)$ is infinite, $\widehat{\partial} f(x) = \partial f(x) = \emptyset$.

While the regular subgradient seems simpler and more appealing at first, we will use the general subgradient in all the following, simply referenced under the name subgradient for simplicity. The reason for this is its superior continuity properties as stated in the following lemma.

Lemma A.2 (Rockafellar and Wets 2009, Th. 8.6, Prop. 8.7 [⋆])**.** *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which $f(x)$ is finite, then the sets of regular subgradients $\widehat{\partial} f(x)$ and general subgradients $\partial f(x)$ are closed. Furthermore, the set of general subgradients ∂f is outer semi-continuous at x, ie.*

$$\limsup_{u \to x \text{ with } f(u) \to f(x)} \partial f(u) := \{v : \exists u_r \to x, \exists\, v_r \to v \text{ with } v_r \in \partial f(u_r)\} \subset \partial f(x)$$

Note that the regular and limiting subdifferentials at some point x coincide in a variety of situations, we then say that the function is *(Clarke) regular* at x (Rockafellar and Wets, 2009, Def. 7.25, Cor. 8.11). While less natural in its definition, the outer semi-continuity property of the general subgradient allows us, for example, to deduce that any limit point x of a sequence (x_k) satisfy $0 \in \partial f(x)$ if the distance from $\partial f(x_k)$ to 0 vanishes.

The condition $0 \in \partial f(x)$ is particularly interesting since it is related to local minimas by Fermat's rule.

Theorem A.3 (Fermat's rule)**.** *If a proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ has a local minimum at x (ie. if there is a neighborhood $\mathcal{U}$ of x such that $f(x) \leq f(u)$ for all $u \in \mathcal{U}$) then $0 \in \partial f(x)$.*

A.2 Differentiability

Differentiability plays a central role in optimization. This is somehow a special case of the notion of subgradient defined above but the treatment of differentiable functions will be rather different algorithmically. In order to promote even more this difference, we will adopt the following convention for the name of generic functions: (*i*) f if it is differentiable; (*ii*) g if it is not assumed differentiable; and (*iii*) f if the differentiability does not play a role in the result.

A.2.1 Derivative of a function from $\mathbb{R}$ to $\mathbb{R}$

In this basic case, the notion of differentiability is quite direct.

Definition A.4. A function $f : \mathcal{V} \subset \mathbb{R} \to \mathbb{R}$ defined on a open subset[24] $\mathcal{V}$ of $\mathbb{R}$ is differentiable at $x \in \mathcal{V}$ if the derivative (ie. the limit)

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

exists. This function f is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$.

[24] At first read, you can take $\mathcal{V}$ as the full space to fix ideas

This definition is equivalent to the existence of a real number $f'(x)$ such that

$$f(x+h) = f(x) + f'(x)h + o(|h|).$$

Note that we now only consider an open subset of $\mathbb{R}$ over which the function is finite-valued. If f takes infinite values on any open set containing x, then it cannot be differentiable at that point.

In addition, if f is differentiable at x, it is necessarily continuous at x. The derivative f' is itself a function from $\mathcal{V} \to \mathbb{R}$ and may also be continuous (on $\mathcal{V}$), in which case, we say that f is continuously differentiable, often denoted $C^1(\mathcal{V})$ or simply C^1.

The derivative of the derivative is called the second-order derivative, noted f''. If it exists and is continuous, we say that f is C^2. Iterating, we can easily define higher order derivatives and differentiability classes up to C^∞.

A.2.2 Gradient of a function from $\mathbb{R}^n$ to $\mathbb{R}$

Let us now consider a function defined over an open subset $\mathcal{V}$ of $\mathbb{R}^n$

$$\begin{array}{rccc} f: & \mathcal{V} \subset \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x = [x_1, .., x_n] & \longmapsto & f(x) \end{array}.$$

For every $x \in \mathcal{V}$, the i-th *partial function* is defined on $\mathcal{V}' \subset \mathbb{R}$ as

$$\begin{array}{rccc} \phi_{i,x}: & \mathcal{V}' & \longrightarrow & \mathbb{R} \\ & u & \longmapsto & f(x_1, .., x_{i-1}, u, x_{i+1}, .., x_n) \end{array},$$

and since this function falls into the case of the previous section, we can study its differentiability. If for all i, $\phi_{i,x}$ is differentiable at x_i, then, we will say that f is differentiable at x.

Definition A.5. A function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^n$ is differentiable at $x \in \mathcal{V}$ if for all $i = 1, .., n$, the derivative (ie. the limit)

$$\frac{\partial f}{\partial x_i}(x) := \lim_{h \to 0} \frac{\phi_{i,x}(x_i + h) - \phi_{i,x}(x_i)}{h}$$

exists. This function f is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$. Further, if f is differentiable on $\mathcal{V}$, we define its *gradient* as the $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}^n$ mapping

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

Similar to what was obtained in the one-dimensional case, we have a *first-order* development of f at a point x at which f is differentiable:

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

A.2.3 Jacobian of a mapping $\mathbb{R}^m$ to $\mathbb{R}^n$

Now, let us consider the case of a mapping (ie. a multi-valued function) from $\mathbb{R}^m$ to $\mathbb{R}^n$

$$\begin{array}{rccc} c: & \mathcal{V} \subset \mathbb{R}^m & \longrightarrow & \mathbb{R}^n \\ & x = [x_1, .., x_m] & \longmapsto & c(x) = [c_1(x), .., c_n(x)] \end{array}.$$

A mapping is differentiable if and only if each of its *component functions* is differentiable as formalized in the following definition.

Definition A.6. A mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^m$ is differentiable at $x \in \mathcal{V}$ if for all $i = 1, .., n$, and all $j \in 1, .., m$, the derivative $\frac{\partial c_i}{\partial x_j}(x)$ exists. This mapping c is differentiable on $\mathcal{V}$ if it is differentiable at every point of $\mathcal{V}$. Further, if c is differentiable on $\mathcal{V}$, we define its *Jacobian* as the $\mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ mapping[25]

[25] The name comes from Carl Gustav Jacob Jacobi (1804-1851), a German mathematician.

$$Jc(x) = \begin{bmatrix} \nabla c_1(x)^\top \\ \vdots \\ \nabla c_n(x)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial c_1}{\partial x_1}(x) & \dots & \frac{\partial c_1}{\partial x_m}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial c_n}{\partial x_1}(x) & \dots & \frac{\partial c_n}{\partial x_m}(x) \end{bmatrix}.$$

While, we do not often differentiate mappings, we often differentiate compositions of a function and mapping. For this, the *chain rule* gives a efficient formula based on the respective gradients and Jacobian of the functions.

Lemma A.7 (Chain rule)**.** *Take a function $f : \mathcal{V}' \subset \mathbb{R}^n \to \mathbb{R}$ and a mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$. If c is differentiable at $x \in \mathcal{V}$ and f is differentiable at $c(x) \in \mathcal{V}'$, then $f \circ c$ is differentiable at x and its gradient can be obtained by*[26]

[26] $f \circ c(x) = f(c(x))$

$$\nabla f \circ c(x) = Jc(x)^\top \nabla f(c(x)). \qquad \text{(Chain rule)}$$

The first-order development of $f \circ c$ is thus

$$f \circ c(x + h) = f \circ c(x) + \langle Jc(x)^\top \nabla f(c(x)), h \rangle + o(\|h\|).$$

A.2.4 Second-order differentiability

The derivative of the gradient, that is the second-order derivative of the function, is often used in numerical optimization methods.

Definition A.8. A function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}$ is twice differentiable at $x \in \mathcal{V}$ if its gradient is differentiable at $x \in \mathcal{V}$.
Further, if f is twice differentiable on $\mathcal{V}$, we define its *Hessian* as the $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ mapping[27]

[27] also denoted by Hf, its name comes from Ludwig Otto Hesse (1811-1874), a German mathematician.

$$\nabla^2 f(x) = J\nabla f(x) = \begin{bmatrix} \frac{\partial^2 f}{(\partial x_1)^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \dots & \frac{\partial^2 f}{(\partial x_n)^2}(x) \end{bmatrix}.$$

This definition comes with the following important property.

Lemma A.9. *The Hessian of a function $f : \mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathcal{V}$ is a symmetric matrix.*

Proof. This follows directly from Schwarz's theorem.[28] □

[28] Hermann Schwarz (1843-1921), German mathematician, was the first to propose a rigorous proof of the symmetry of second derivatives (also called the equality of mixed partials).

Remark A.10 (Hessian at a local minimum). If f admits a local minimum at x and is twice differentiable at x, then $\nabla f = 0$ by Fermat's rule (Theorem A.3) but we can also show that $\nabla^2 f(x)$ is positive semi-definite; see **?? ??**. ◀

A.2.5 Fréchet derivatives

The notion of Fréchet derivatives generalizes the notion of gradient and Jacobian seen above. A mapping $c : \mathcal{V} \subset \mathbb{R}^m \to \mathbb{R}^n$ defined on a open subset $\mathcal{V}$ of $\mathbb{R}^m$ is *Fréchet differentiable* at $x \in \mathcal{V}$ if there exists a linear operator

$$\begin{array}{rccc} \mathrm{D}c(x) : & \mathbb{R}^m & \longrightarrow & \mathbb{R}^n \\ & h & \longmapsto & \mathrm{D}c(x)[h] \end{array}$$

called the (Fréchet) *differential* of c at x,[29] such that

$$c(x+h) = c(x) + \mathrm{D}c(x)[h] + o(\|h\|)$$
$$\text{or, equivalently } \lim_{h \to 0} \frac{\|c(x+h) - c(x) - \mathrm{D}c(x)[h]\|}{\|h\|} = 0.$$

[29] from Maurice René Fréchet (1878-1973), a French mathematician.

Then, if f is a $\mathcal{V} \subset \mathbb{R}^n \to \mathbb{R}$ function, the gradient of f can be defined as the unique element of $\mathbb{R}^n$ that satisfies

$$\mathrm{D}f(x)[h] = \langle \nabla f(x), h \rangle \text{ for all } h \in \mathbb{R}^n$$

and thus, in line with the regular subgradient notation, it can also be defined as

$$\nabla f(x) = \{v : f(u) = f(x) + \langle v, u - x \rangle + o(\|u - x\|) \text{ for all } u \in \mathbb{R}^n\}. \tag{A.3}$$

The same can be done for mappings and the Jacobian of c can be defined as the unique $\mathbb{R}^n \times \mathbb{R}^m$ operator $Jc(x)$ such that $\mathrm{D}c(x)[h] = Jc(x)h$.

Finally, the Chain rule for differentials is

$$\mathrm{D}(f \circ c)(x)[h] = \mathrm{D}f(c(x))[\mathrm{D}c(x)[h]] = \langle \nabla f(c(x)), Jc(x)h \rangle = \langle Jc(x)^\top \nabla f(c(x)), h \rangle.$$

A.2.6 Link with subdifferentials

To be complete, let us relate the notion of gradient defined here with the subdifferentials defined before.

Lemma A.11. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \mathbb{R}^n$ at which f is differentiable, then $\nabla f(x) = \widehat{\partial} f(x) \subset \partial f(x)$. If, in addition, f is continuously differentiable around x, then $\nabla f(x) = \partial f(x)$.*

Proof. For the first part, interpret directly (A.3) as (A.1). For the second part, the continuity of ∇f enables leaves no other choice for a limit in (A.2) than $\nabla f(x)$. □

In the common case, where we deal with the sum of two functions, the following lemma is particularly useful.

Lemma A.12. *If $F = f + g$ with f continuously differentiable around x and $g(x)$ finite, then $\partial F(x) = \nabla f(x) + \partial g(x)$.*

Proof. Direct from the definitions. □

A.3 Smoothness and Gradient descent

There is slight discrepancy in the literature concerning the notion of smoothness for functions. In (Rockafellar and Wets, 1998), it is used for continuously differentiable functions, in Riemannian analysis it often refers to C^∞ function, while in numerical optimization and machine learning (see eg. (Bubeck et al., 2015)), it is used for functions with Lipschitz-continuous gradients. We will adopt the latter viewpoint. The reason for this is that it allows us to have a quadratic upper approximation of our function, obtained directly from the fundamental theorem of calculus. This is the crucial point for the use of gradient methods.

Definition A.13. We say that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is L-smooth if it has a L-Lipschitz continuous gradient, ie. if

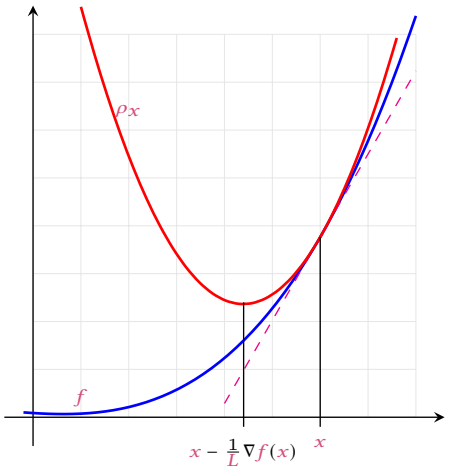
$$\|\nabla f(x) - \nabla f(u)\| \leq L\|x - u\| \text{ for all } x, u \in \mathbb{R}^n.$$

From this property, we can derive this highly important lemma.

Lemma A.14. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with a L-Lipschitz continuous gradient, then for any $x, u \in \mathbb{R}^n$, one has*

$$|f(u) - f(x) - \langle \nabla f(x), u - x\rangle| \leq \frac{L}{2}\|x - u\|^2.$$

Thus, if we fix a point x, the function $\rho_x : u \mapsto f(x) + \langle \nabla f(x), u - x\rangle + \frac{L}{2}\|u - x\|^2$ is quadratic in its argument and majorizes f, that is to say $\rho_x(u) \geq f(u)$ for any u. Furthermore, the minimum of ρ_x is attained at $x^\star = x - \frac{1}{L}\nabla f(x)$.



Such a quadratic approximation can be leveraged using gradients steps, ie. taking

$$u = x - \gamma \nabla f(x)$$

for some $\gamma > 0$. Indeed, in that case, Lemma A.14 gives us

$$f(u) \leq f(x) - \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x - u\|^2 = f(x) - \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla f(x)\|^2. \tag{A.4}$$

Thus, taking a gradient step leads to a strict functional decrease ($f(u) < f(x)$) as soon as $\gamma < 2/L$. This is the core idea behind the *gradient descent* algorithm.[30] Take $x_0 \in \mathbb{R}^n$ and $\gamma > 0$, the gradient descent algorithm consists in iterating

$$x_{k+1} = x_k - \gamma \nabla f(x_k) \qquad \text{(Gradient descent)}$$

and leads to the following guarantees.

[30] introduced by Louis Augustin Cauchy (1789–1857), a French mathematician, in his "Compte Rendu à l'Académie des Sciences" of October 18, 1847.

Theorem A.15. *Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with a L-Lipschitz continuous gradient and such that $\inf f > -\infty$. Assume that (Gradient descent) is run with $0 < \gamma < 2/L$, then $(f(x_k))$ converges and any limit point $\overline{x}$ of (x_k) satisfies $\nabla f(\overline{x}) = 0$.*

Even though the above theorem is only a partial justification, gradient descent is widely used for finding critical points of smooth functions. The link between finding critical points and minimizing a function will be brought in the next chapter by convexity. In that case, the guarantees of gradient descent will be strengthened.

Finally, let us conclude this part with a quote from the original paper by Cauchy (Cauchy et al., 1847) that also applies to us *"I'll restrict myself here to outlining the principles underlying [my method], with the intention to come again over the same subject"*.[31]

[31] In the original text: *" Je me bornerai pour l'instant à indiquer les principes sur lesquels [ma méthode] se fonde, me proposant de revenir avec plus de détails sur le même sujet, dans un prochain mémoire."*. The translation and reference is due to Claude Lemaréchal, see (Lemaréchal, 2012).

Remark A.16 (What if my differentiable function is not smooth [⋆]). If f is differentiable but not smooth, these guarantees fall down. We have to take a closer look at the function:

- if the function seems locally Lipschitz but not constant can be computed, then you can numerically test different values and see if (A.4) is satisfied (see later);
- if the function is blowing up at some finite point, a change of geometry may help (see the Operation Research complementary);
- otherwise, treat it as a non-smooth function.

◀

APPENDIX B CONVEXITY AND OPTIMALITY

CONVEXITY is at the heart of optimization. This is notably due to the unicity of projections onto convex sets and the direct link between critical points and minimums for convex functions.

In this chapter, we will first study convex *sets*, then convex *functions*.

B.1 CONVEX SETS

B.1.1 Motivation: Projecting onto a closed set

Similarly to orthogonal projections onto affine subspaces, we can define projection on *nonempty closed* sets.[32]

Thus, let us consider a non-empty closed set C and investigate the problem

$$\inf_{x\in C} f_y(x) := \frac{1}{2}\|y-x\|^2 \tag{B.1}$$

which intuitively amounts to projecting y onto C.

First, take $u \in C$, and define $S := \{x \in \mathbb{R}^n : \|y-x\|^2 \leq \|y-u\|^2\}$. Then, the problem (B.1) is equivalent to

$$\inf_{x\in C\cap S} f_y(x) := \frac{1}{2}\|y-x\|^2 \tag{B.2}$$

where $C \cap S$ is a closed compact set. Projecting thus amounts to minimizing a continuous function over a closed compact set, which always admits a solution, as per the following lemma.

Lemma B.1. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semi-continuous function (or in particular, a continuous function) and let S be a closed compact set. Then, there is some $x^\star \in S$ such that $f(x^\star) = \inf_{x\in S} f(x)$.*

Proof. ([⋆]) Since f is proper, it nevers takes the value $-\infty$ thus $\bar{\beta} := \inf_{x\in S} f(x) > -\infty$. For a decreasing sequence of reals (β_n) with $\beta_n \to \bar{\beta}$, let us define the sequence of the $S_{\beta_n} = \{x : f(x) \leq \beta_n\}$. For any n, S_{β_n} is nonempty, closed, and included in $S_{\beta_{n-1}}$. Thus, the limit $S_{\bar{\beta}} = \{x : f(x) = \inf_{u\in S} f(u)\}$ is also nonempty and closed which gives the result. □

This grants the existence of a minimizer of (B.2), and thus of (B.1), ie. a projection on C. In particular, the inf above are actually min. However, the projection may not be unique, that is where convexity comes into play.[33]

[32] Nonempty: otherwise there is nothing to project onto. Closed: otherwise "the" closest point in a set from another point is not well-defined.

[33] The above enables us to show the existence of projections onto nonempty closed sets, but the projection may not be unique.

B.1.2 Convexity for sets

Let us now introduce the definition of a convex set.

Definition B.2. A subset C of $\mathbb{R}^n$ is convex if and only if for any $x, u \in C$, $(1-\alpha)x + \alpha u \in C$ for any $\alpha \in (0,1)$.

The crucial property here is that any (weighted) average of points of a convex set belongs stay in the set. Equivalently, the set C is convex if and only if for any $(x_1, .., x_N) \in C^N$,

$$\sum_{i=1}^{N} \alpha_i x_i \in C \text{ for any } (\alpha_1, .., \alpha_N) \in \mathbb{R}_+^N \text{ with } \sum_{i=1}^{N} \alpha_i = 1,$$

where $\sum_{i=1}^N \alpha_i x_i$ is called a *convex combination* of $(x_1, .., x_N)$.

Examples of convex sets:

- Affine spaces $\{x : \langle s, x\rangle = r\}$
- Balls $\{x : \|x - s\| \leq r\}$
- Half spaces $\{x : \langle s, x\rangle \leq r\}$ and open half spaces $\{x : \langle s, x\rangle < r\}$
- Simplices $\{x : \sum_{i=1}^n x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i = 1, .., n\}$
- Intersections of convex sets $\cap_{i=1}^N C_i$

Examples of non-convex sets:

- Discrete sets (eg. $\{0\} \cup \{1\}$) or disjoint sets
- Spheres $\{x : \|x - s\| = r\}$
- Sets with "holes"

B.1.3 Projection on convex sets

Getting back to the projection problem (B.1)

$$\min_{x \in C} f_y(x) := \frac{1}{2}\|y - x\|^2 \tag{B.3}$$

where $S := \{x \in \mathbb{R}^n : \|y - x\|^2 \leq \|y - u\|^2\}$. Now, let us assume that C is additionally convex.

Suppose that ${x_1}^\star \neq {x_2}^\star$ are two distinct solutions of (B.3). Define ${x_0}^\star = ({x_1}^\star + {x_2}^\star)/2$, then

$$\begin{aligned} f_y({x_0}^\star) &= \frac{1}{2}\|y - {x_0}^\star\|^2 = \frac{1}{2}\|(y - {x_1}^\star)/2 + (y - {x_2}^\star)/2\|^2 \\ &= \frac{1}{4}\|y - {x_1}^\star\|^2 + \frac{1}{4}\|y - {x_2}^\star\|^2 - \frac{1}{8}\|{x_1}^\star - {x_2}^\star\|^2 \\ &= \frac{1}{2}(f_y({x_1}^\star) + f_y({x_2}^\star)) - \frac{1}{8}\|{x_1}^\star - {x_2}^\star\|^2 \end{aligned}$$

thus $f_y({x_0}^\star) < f_y({x_1}^\star) = f_y({x_2}^\star)$ which contradicts ${x_1}^\star \neq {x_2}^\star$ being two distinct solutions. Hence, the projection on a convex set is unique. We have shown the following lemma.

Lemma B.3. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, there is a unique projection $\mathrm{proj}_C(y)$, solution of (B.3).*

In fact, this unique projection can be characterized more precisely.

Theorem B.4. *Let C be a closed nonempty convex set. Then, for any $y \in \mathbb{R}^n$, $\mathrm{proj}_C(y)$ is the projection of y onto C if and only if*

$$\langle y - \mathrm{proj}_C(y), z - \mathrm{proj}_C(y) \rangle \leq 0 \text{ for all } z \in C.$$

Proof. Left as an exercise. See (Hiriart-Urruty and Lemaréchal, 1993b, Th. 3.1.1). □

B.1.4 Minimization over convex sets

Now, let us consider a more general problem: minimizing a function f over a convex set C. The problem consists in finding $x^\star \in C$ such that $f(x^\star) \leq f(x)$ for all $x \in C$, we note this problem

$$x^\star \in \mathrm{argmin}_C f \Leftrightarrow x^\star \text{ is a solution of } \inf_{x \in C} f(x)$$

We directly note that if C is empty, the problem is impossible[34] and if C is open it may be impossible to find a solution. Hence, we will restrict our analysis to closed nonempty convex sets as before.

[34] *infeasible* in the optimization language.

Here, a helpful notion is the one of normal cone.

Definition B.5 (Normal cone)**.** The *normal cone* of a convex set C at a point $x \in C$ is defined as the set $N_C(x) := \{u : \langle y - x, u \rangle \leq 0 \text{ for all } y \in C\}$.

The *constrained* variant of Fermat's rule (Theorem A.3) that links the (sub)gradient of the function with local minimas writes as follows.

Theorem B.6 ((Rockafellar and Wets, 1998, Th. 6.12,8.15))**.** *If a proper lower-semicontinuous function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ has a local minimum at x constrained to the convex set C (ie. if there is a neighborhood $\mathcal{U}$ of x in C such that $f(x) \leq f(u)$ for all $u \in \mathcal{U}$) then $0 \in \partial f(x) + N_C(x)$ or, equivalently,*

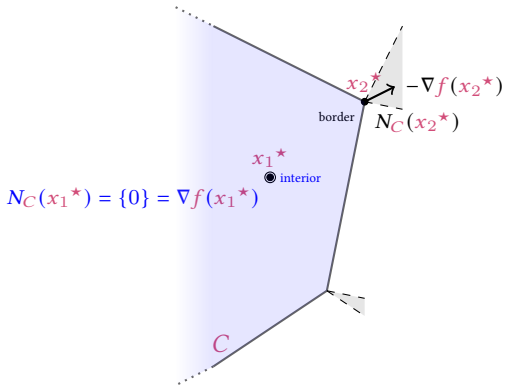
$$\langle y - x, v \rangle \geq 0$$

for any $v \in \partial f(x)$ and all $y \in C$.
In particular, if f is differentiable, $0 \in \nabla f(x) + N_C(x)$ means that

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

for all $y \in C$.

Note that if x belongs to the relative interior of C, then $N_C(x) = \{0\}$.



B.2 Convex functions

The notion of convexity is as important for functions as for sets. Notably, this is the notion that will enable us to go from the (sub)gradient inequalities and local minimizers above to *global* minimizers.

B.2.1 Definition

[35] This is the set $\mathrm{epi} f := \{(x,t) : f(x) \leq t\}$

A function is convex if and only if its *epigraph*[35] is convex. However, the following definition is much more direct.

Definition B.7. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if for any $x, u \in \mathrm{dom}\, f$, $f((1-\alpha)x + \alpha u) \leq (1-\alpha)f(x) + \alpha f(u)$ for any $\alpha \in (0,1)$.

More generally convex functions verify *Jensen's inequality*. For any convex combination $\sum_{i=1}^N \alpha_i x_i$,

$$f\left(\sum_{i=1}^N \alpha_i x_i\right) \leq \sum_{i=1}^N \alpha_i f(x_i).$$

Checking the definition directly may be possible but it is often simpler to rely on convexity-preserving operations :

- all norms are convex;
- a sum of convex functions is convex;
- affine substitution of the argument (if f is convex, $x \mapsto f(Ax+b)$ is convex for any affine map $Ax+b$);
- the (pointwise) maximum of convex functions is convex.

The most striking point of convex functions is that local minimizers are actually global.

Theorem B.8. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. Then, every local minimizer of f is a (global) minimizer.*

B.2.2 Subgradients of convex functions

This class of functions comes with several interesting properties, for instance $\mathrm{dom}\, f$ and $\mathrm{argmin}\, f$ are convex if f is convex, furthermore, every local minimum is a global

one. This is again captured by the notion of subgradients.

Lemma B.9 (Rockafellar and Wets 1998, Prop. 8.12). *Consider a convex proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \operatorname{dom} f$. Then,*

$$\partial f(x) = \{v : f(u) \geq f(x) + \langle v, u - x\rangle \text{ for all } u \in \mathbb{R}^n\} = \widehat{\partial} f(x) \neq \emptyset.$$

Thus, f is regular at any point and $0 \in \partial f(x)$ if and only if $x \in \operatorname{argmin} f$.

An important point is that $u \mapsto f(x)+\langle v, u-x\rangle$ provides a linear under-approximation of the whole function f.

Furthermore, we have the same link between subgradients and optimality when constrained to a convex set.

Theorem B.10 ((Rockafellar and Wets, 1998, Th. 8.15)). *Consider a proper lower-semicontinuous convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a convex set C. Then, $x \in \operatorname{argmin}_C f$ if and only if $0 \in \partial f(x) + N_C(x)$ or, equivalently,*

$$\langle y - x, v\rangle \geq 0$$

for any $v \in \partial f(x)$ and all $y \in C$.

B.2.3 Differentiable convex functions

First, Theorem B.10 can be a little simplified if the function is differentiable.

Theorem B.11 ((Rockafellar and Wets, 1998, Th. 6.12)). *Consider a proper lower-semicontinuous convex and differentiable function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a convex set C. Then, $x \in \operatorname{argmin}_C f$ if and only if $0 \in \nabla f(x) + N_C(x)$ which means that*

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

for all $y \in C$.

In addition, for a differentiable f, convexity can be seen directly as a property on the gradient mapping.

Theorem B.12 (Bauschke and Combettes 2011, Prop. 17.10). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function with open domain.[36] Suppose that f is differentiable on $\operatorname{dom} f$. Then the following are equivalent:*

i) f is convex;

ii) $f(u) \geq f(x) + \langle \nabla f(x), u - x\rangle$ for all $x, u \in \operatorname{dom} f$;

iii) $\langle \nabla f(x) - \nabla f(u), x - u\rangle \geq 0$ for all $x, u \in \operatorname{dom} f$, ie. ∇f is monotone.

Furthermore, if f is twice differentiable on $\operatorname{dom} f$, any of the above is equivalent to

iv) $\langle u, \nabla^2 f(x) u\rangle \geq 0$ for all $x, u \in \operatorname{dom} f$, ie. $\nabla^2 f$ is positive semi-definite.

[36] *typically here, $\operatorname{dom} f = \mathbb{R}^n$.*

B.2.4 Strict convexity

Strict convexity is simply convexity but when every inequality is replaced with a *strict inequality*: a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strictly convex if and only if for any $x, u \in C$, $f((1-\alpha)x + \alpha u) < (1-\alpha)f(x) + \alpha f(u)$ for any $\alpha \in (0, 1)$. All results above then hold with strict inequalities.

Lemma B.13. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strictly convex lower semi-continuous proper function and C a convex set, then f has at most one minimizer on C. In particular, f has at most one minimizer on $\mathbb{R}^n$.*

Strict convexity can be observed mathematically and from that we can ensure the uniqueness of solutions. However, it is almost impossible to exploit numerically since it only grants us a strict inequality and not an exploitable knowledge about the function's local behavior. For this, we need a stronger condition: strong convexity.

B.2.5 Strong convexity

While convexity provides affine lower bounds, having quadratic lower-bounds enable to get a better control that may have a great impact on the convergence of optimization methods; this is captured by the notion of strong convexity.

Definition B.14. For some $\mu > 0$, a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is μ-strongly convex if and only if $f - \frac{1}{2}\mu\|\cdot\|^2$ is convex.

Using the fact that $g := f - \frac{1}{2}\mu\|\cdot\|^2$ is convex and verifies $\partial g = \partial f - \mu\cdot$ by Lemma A.12, we get that for any $x \in \mathbb{R}^n$ and any $v \in \partial f(x)$

$$f(u) \geq f(x) + \langle v, u - x\rangle + \frac{\mu}{2}\|u - x\|^2 \text{ for all } u \in \mathbb{R}^n \tag{B.4}$$

which directly implies that a strongly convex function has at most one minimizer by taking x such that $0 \in \partial f(x)$. The following lemma then adds the existence (see (Bauschke and Combettes, 2011, Chap. 11.4) for a more general take).

Lemma B.15. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strongly convex lower semi-continuous proper function and C a convex set, then f has exactly one minimizer on C. In particular, f has exactly one minimizer one minimizer on $\mathbb{R}^n$.*

Proof. ([⋆]) Let us consider the case where $C = \mathbb{R}^n$, the other cases can be deduced easily. From (B.4), we get that for all $u \in \mathbb{R}^n$,

$$\begin{aligned} f(u) &\geq f(x) + \frac{\mu}{2}\|x\|^2 - \langle v, x\rangle + \langle v + \mu x, u\rangle + \frac{\mu}{2}\|u\|^2 \\ &\geq f(x) + \frac{\mu}{2}\|x\|^2 - \langle v, x\rangle - \|v + \mu x\|\|u\| + \frac{\mu}{2}\|u\|^2 \end{aligned}$$

hence $f(u)/\|u\| \to +\infty$ when $\|u\| \to +\infty$, ie. f is supercoercive. Thus, this means that for any t, the level set $\{x : f(x) \leq t\}$ is bounded (this is direct by contradiction, see (Bauschke and Combettes, 2011, Chap. 11.11)). This means that since f is proper, we can take t sufficiently large so that the corresponding level set is non-empty and bounded. Finally, since f is lower semi-continuous, applying Lemma B.1 to this compact set gives us the existence of a minimal value, which is unique from the quadratic lower bound expressed in (B.4). □

If a differentiable function is strongly convex, we have the following characterizations.

Theorem B.16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function with open domain. Suppose that f is differentiable on $\operatorname{dom} f$. Then the following are equivalent:*

i) f is μ-strongly convex;

ii) $f(u) \geq f(x) + \langle \nabla f(x), u - x\rangle + \frac{\mu}{2}\|u - x\|^2$ for all $x, u \in \operatorname{dom} f$;

iii) $\langle \nabla f(x) - \nabla f(u), x - u\rangle \geq \mu\|u - x\|^2$ for all $x, u \in \operatorname{dom} f$, ie. ∇f is monotone.

Furthermore, if f is twice differentiable on $\operatorname{dom} f$, any of the above is equivalent to

iv) $\langle u, \nabla^2 f(x) u \rangle \geq \mu \|u\|^2$ for all $x, u \in \operatorname{dom} f$, ie. $\nabla^2 f$ is positive definite.

APPENDIX C DUALITY BETWEEN MEASURES AND FUNCTIONS

DUALITY between measures and functions is at the core of several results of the course.

C.1 Duality Between Measures and Functions on Subsets of $\mathbb{R}^d$

C.1.1 Setting

Let $X \subset \mathbb{R}^d$ be a (not necessarily bounded) subset, endowed with the topology induced by the Euclidean metric. Then:

- X is locally compact,
- X is Hausdorff,
- the Borel σ-algebra on X is well defined.

These properties follow automatically from the fact that $\mathbb{R}^d$ is a locally compact Hausdorff space.

Definition C.1 (Positive Radon Measure)**.** A *positive Radon measure* on X is a map

$$\mu : \mathcal{B}(X) \to [0, +\infty]$$

satisfying:

1. **Countable additivity:**
$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n) \quad \text{for disjoint Borel sets } E_n.$$
2. **Local finiteness:**
$$\mu(K) < \infty \quad \text{for all compact } K \subset X.$$
3. **Inner regularity:**
$$\mu(E) = \sup\{\mu(K) : K \subset E,\ K \text{ compact}\}.$$
4. **Outer regularity:**
$$\mu(E) = \inf\{\mu(U) : E \subset U,\ U \text{ open}\}.$$

We denote by $\mathcal{M}^+(X)$ the set of all positive Radon measures on X.

Remark. In many functional-analytic statements (including the Riesz representation theorem), one encounters *signed* Radon measures. These arise automatically as differences of positive ones:

$$\mu = \mu^+ - \mu^-.$$

However, the fundamental object is the positive Radon measure, and it suffices to define it first.

Let $C(X)$ be the space of continuous functions on X and

$$C_0(X) = \{f \in C(X) : f(x) \to 0 \text{ as } |x| \to \infty\}.$$

Equipped with the supremum norm,

$$\|f\|_\infty = \sup_{x \in X} |f(x)|,$$

this is a Banach space.

C.1.2 Duality via Integration

Every positive Radon measure μ defines a linear functional on $C_0(X)$ by

$$T_\mu(f) := \int_X f(x)\, d\mu(x).$$

This map is continuous and satisfies

$$\|T_\mu\| = \mu(X).$$

C.1.3 Riesz Representation Theorem

Theorem C.2 (Riesz–Markov–Kakutani). *Let $X \subset \mathbb{R}^d$. Then:*

$$(C_0(X))^* \cong \mathcal{M}(X),$$

where:

- *$\mathcal{M}(X)$ is the space of finite signed Radon measures,*
- *the identification is given by*

$$T(f) = \int_X f\, d\mu,$$

- *every continuous linear functional arises uniquely in this way.*

C.1.4 Meaning of the Symbol $\cong$

The notation

$$(C_0(X))^* \cong \mathcal{M}(X)$$

means that:

- the two spaces are *linearly isomorphic*,

- the correspondence is *isometric*:

$$\|T\|_{(C_0)^*} = \|\mu\|_{\mathcal{M}},$$

- the isomorphism is *canonical*, given by integration.

Thus $\cong$ denotes an *isometric isomorphism of Banach spaces*, not merely a bijection.

C.1.5 Interpretation

This duality expresses the fundamental principle:

> "Measures are continuous linear functionals on spaces of test functions."

It underlies:

- weak convergence of measures,
- probability theory on $\mathbb{R}^d$,
- weak formulations of PDEs,
- distribution theory.

C.2 The Dual of the Space of Radon Measures

C.2.1 Setting

Let $X \subset \mathbb{R}^d$, and let

$$\mathcal{M}(X)$$

denote the Banach space of finite signed Radon measures on X, equipped with the total variation norm

$$\|\mu\|_{\mathcal{M}} = |\mu|(X).$$

Recall from the Riesz–Markov–Kakutani theorem that

$$\mathcal{M}(X) \cong (C_0(X))^*.$$

We now ask the *reverse question*:

> What is the (topological) dual of $\mathcal{M}(X)$?

C.2.2 The Canonical Embedding into the Double Dual

Since $\mathcal{M}(X)$ is a Banach space, its dual is

$$\mathcal{M}(X)^* = (C_0(X))^{**}.$$

There is always a canonical isometric embedding

$$C_0(X) \hookrightarrow (C_0(X))^{**} = \mathcal{M}(X)^*,$$

given by

$$f \mapsto \left(\mu \mapsto \int_X f \, d\mu\right).$$

Thus every function $f \in C_0(X)$ defines a continuous linear functional on $\mathcal{M}(X)$ via

$$\Phi_f(\mu) := \int_X f \, d\mu.$$

C.2.3 Identification with Bounded Continuous Functions

If X is a locally compact subset of $\mathbb{R}^d$, then:

$$\boxed{\mathcal{M}(X)^* \cong C_b(X)}$$

where:

- $C_b(X)$ is the space of bounded continuous functions,
- the pairing is
$$\langle f, \mu \rangle = \int_X f \, d\mu,$$
- the norm is $\|f\|_\infty$.

This identification holds in the following precise sense:

- Every $f \in C_b(X)$ defines a bounded linear functional on $\mathcal{M}(X)$.
- The map
$$C_b(X) \to \mathcal{M}(X)^*$$
is an isometric embedding.
- If X is compact, this map is onto.
- If X is not compact, the inclusion is strict:

$$C_0(X) \subsetneq C_b(X) \subsetneq \mathcal{M}(X)^*.$$

C.2.4 Interpretation

The duality chain is therefore:

$$C_0(X) \xrightarrow{\text{dual}} \mathcal{M}(X) \xrightarrow{\text{dual}} C_b(X)$$

with:

$$f \in C_b(X) \quad \leftrightarrow \quad \left(\mu \mapsto \int_X f \, d\mu\right).$$

This expresses the hierarchy:

compactly vanishing functions $\subset$ bounded continuous functions $\subset$ all linear functionals on measures.

C.2.5 Important Remark (Why Not All of $(C_0)^{**}$?)

Although

$$\mathcal{M}(X)^* = (C_0(X))^{**},$$

the canonical embedding

$$C_0(X) \hookrightarrow (C_0(X))^{**}$$

is not surjective unless $C_0(X)$ is reflexive, which happens only in trivial cases.

Thus:

- $\mathcal{M}(X)^*$ is strictly larger than $C_0(X)$,
- $C_b(X)$ provides a concrete realization of this dual,
- additional elements of $(C_0)^{**}$ correspond to highly nonlocal objects.

C.2.6 Summary Diagram

$$\boxed{C_0(X) \subset C_b(X) \cong \mathcal{M}(X)^*}$$

$$\boxed{(C_0(X))^* = \mathcal{M}(X)}$$

$$\boxed{(C_0(X))^{**} = \mathcal{M}(X)^*}$$

C.2.7 Conceptual Interpretation

- Measures act on functions by integration.
- Functions act on measures by evaluation.
- The duality is symmetric but not reflexive.
- Noncompactness of X is responsible for the strict inclusions.

C.3 The Case of Compact $X \subset \mathbb{R}^d$

C.3.1 Simplifications Due to Compactness

Assume now that

$$X \subset \mathbb{R}^d \quad \text{is compact.}$$

Then several important simplifications occur:

- Every continuous function is automatically bounded.
- $C_0(X) = C(X)$.
- Every Radon measure on X is finite.
- The duality theory becomes exact (no pathologies).

C.3.2 Measures and Functions

Let:

- $C(X)$ = continuous real-valued functions on X,
- $\mathcal{M}(X)$ = finite signed Radon measures on X.

Then:

Theorem C.3 (Riesz–Markov–Kakutani, compact case)**.** *If X is compact, then*

$$(C(X))^* \cong \mathcal{M}(X),$$

isometrically, via

$$T(f) = \int_X f \, d\mu.$$

Thus:

$$\boxed{\text{Measures} \longleftrightarrow \text{linear functionals on } C(X)}$$

C.3.3 The Dual of the Space of Measures

Since $C(X)$ is a Banach space,

$$\mathcal{M}(X)^* = (C(X))^{**}.$$

But now an important simplification occurs:

Theorem C.4. *If X is compact, then*

$$\mathcal{M}(X)^* \cong C(X).$$

C.3.4 Explanation

The identification is given explicitly by:

$$f \in C(X) \quad \longmapsto \quad \Phi_f(\mu) := \int_X f \, d\mu.$$

Moreover:

- Every continuous linear functional on $\mathcal{M}(X)$ is of this form.
- The correspondence is isometric:

$$\|\Phi_f\| = \|f\|_\infty.$$

- No additional pathological elements appear.

Thus:

$$\boxed{\mathcal{M}(X)^* \cong C(X)} \quad \text{(only true when } X \text{ is compact).}$$

C.3.5 Reflexivity Picture

When X is compact:

$$C(X) \xrightarrow{\text{dual}} \mathcal{M}(X) \xrightarrow{\text{dual}} C(X)$$

so the duality closes neatly.
However:

- $C(X)$ is not reflexive unless X is finite,
- but the canonical embedding

$$C(X) \hookrightarrow C(X)^{**}$$

is an isometric isomorphism onto $\mathcal{M}(X)^*$.

C.3.6 Summary Table

Space	Dual	When X is compact
$C(X)$	$\mathcal{M}(X)$	Riesz theorem
$\mathcal{M}(X)$	$C(X)$	Exact duality
$C_0(X)$	$\mathcal{M}(X)$	$C_0(X) = C(X)$

C.3.7 Conceptual Interpretation

When X is compact:

- Measures and continuous functions are in perfect duality.
- No growth or decay conditions are needed.
- Weak-* compactness becomes simple (Banach–Alaoglu).
- This is the natural setting for probability on compact spaces.

> *Compactness removes all analytic pathologies from the duality between functions and measures.*

C.4 Summary Diagram of Dualities

General Case: $X \subset \mathbb{R}^d$ (locally compact)

$$\begin{array}{ccccc} C_0(X) & \xrightarrow{\text{dual}} & \mathcal{M}(X) & \xrightarrow{\text{dual}} & (C_0(X))^{**} \\ \cup & & \| & & \cap \\ C_b(X) & & \mathcal{M}(X) & & \mathcal{M}(X)^* \end{array}$$

$$\boxed{C_0(X) \subsetneq C_b(X) \cong \mathcal{M}(X)^*}$$

Interpretation:

- $C_0(X)^* = \mathcal{M}(X)$ (Riesz theorem)
- $\mathcal{M}(X)^*$ is larger than $C_0(X)$
- $C_b(X)$ embeds isometrically into $\mathcal{M}(X)^*$
- Equality holds only if X is compact

—

Compact Case: $X \subset \mathbb{R}^d$ Compact

When X is compact, we have:

$$C_0(X) = C(X), \qquad C_b(X) = C(X)$$

and the duality collapses to:

$$\boxed{C(X) \xleftrightarrow{\text{dual}} \mathcal{M}(X)}$$

More explicitly:

$$C(X) \xrightarrow{\int f\,d\mu} \mathcal{M}(X) \xrightarrow{\int f\,d\mu} C(X)$$

$$\mathcal{M}(X)^* \cong C(X)$$

—

Conceptual Picture

$$\boxed{\text{Functions} \longleftrightarrow \text{Measures} \longleftrightarrow \text{Functions}} \quad \text{(only when } X \text{ is compact)}$$

- Noncompact X: duality is one-sided
- Compact X: perfect duality
- Failure of compactness = appearance of extra functionals

C.5 Duality: Probabilistic and Functional-Analytic Viewpoints

C.5.1 Probabilistic Version (Probability Measures)

Let $X \subset \mathbb{R}^d$ be compact.

Denote by:

- $\mathcal{P}(X)$ the set of Borel probability measures on X,
- $C(X)$ the space of continuous real-valued functions on X.

Expectation as Duality

Every probability measure $\mu \in \mathcal{P}(X)$ defines a linear functional

$$f \longmapsto \mathbb{E}_\mu[f] := \int_X f \, d\mu.$$

Thus:

$$\mathcal{P}(X) \subset \mathcal{M}(X) = C(X)^*.$$

Weak Convergence

A sequence $(\mu_n) \subset \mathcal{P}(X)$ converges weakly to μ if

$$\int_X f \, d\mu_n \to \int_X f \, d\mu \quad \forall f \in C(X).$$

This is exactly:

$$\mu_n \xrightarrow{w^*} \mu \quad \text{in } \mathcal{M}(X).$$

Compactness (Prokhorov)

Since X is compact:

- $\mathcal{P}(X)$ is tight,
- $\mathcal{P}(X)$ is weakly compact,
- every sequence has a weakly convergent subsequence.

Probabilistic Duality Diagram

$$\boxed{\begin{array}{ccc} C(X) & \xleftrightarrow{\int f\,d\mu} & \mathcal{M}(X) \\ & \supset & \\ & \mathcal{P}(X) & \end{array}}$$

Weak convergence of probabilities $\iff$ weak-* convergence in $C(X)^*$.

—

C.5.2 Functional-Analytic Duality Diagram (with Topologies)

Compact Case

When X is compact:

$$\boxed{\begin{array}{ccc} C(X) & \xleftrightarrow[\text{dual}]{\|\cdot\|_\infty} & \mathcal{M}(X) \\ \| & & \| \\ C(X) & \xleftrightarrow[\text{weak-}^*]{} & C(X)^* \end{array}}$$

Topologies involved:

- $C(X)$: uniform norm topology
- $\mathcal{M}(X)$: total variation norm
- Weak-* topology on $\mathcal{M}(X)$ induced by $C(X)$

—

Non-Compact Case (for comparison)

$$\boxed{\begin{array}{ccccc} C_0(X) & \xrightarrow{\text{dual}} & \mathcal{M}(X) & \xrightarrow{\text{dual}} & (C_0(X))^{**} \\ \cup & & \| & & \cap \\ C_b(X) & & \mathcal{M}(X) & & \mathcal{M}(X)^* \end{array}}$$

Topological Meaning

- $C_0(X)^* = \mathcal{M}(X)$ (Riesz)
- $\mathcal{M}(X)^* \supsetneq C_0(X)$ unless X is compact
- Compactness $\Rightarrow$ no loss of duality
- Noncompactness $\Rightarrow$ extra functionals appear

—

C.5.3 Final Conceptual Summary

$$\boxed{\begin{array}{c}\text{Functions}\\ \updownarrow \\ \text{Measures}\\ \updownarrow \\ \text{Functions}\end{array}} \quad \text{(only when } X \text{ is compact)}$$

- Measures are dual to continuous functions.
- Probability measures are normalized positive elements of this dual.
- Weak convergence is weak-* convergence.
- Compactness makes the duality exact and symmetric.

C.6 L^p Spaces and Duality

C.6.1 Definition of L^p Spaces

Let $X \subset \mathbb{R}^d$ be a measurable set and let λ denote Lebesgue measure. For $1 \le p \le \infty$, define

$$L^p(X) := \{f : X \to \mathbb{R} \text{ measurable } : \|f\|_{L^p} < \infty\},$$

where

$$\|f\|_{L^p} = \begin{cases} \left(\int_X |f(x)|^p \, dx\right)^{1/p}, & 1 \le p < \infty, \\ \operatorname{ess\,sup}_{x \in X} |f(x)|, & p = \infty. \end{cases}$$

Each $L^p(X)$ is a Banach space.

C.6.2 Duality of L^p Spaces

Theorem C.5 (Riesz Representation for L^p). *Let $1 < p < \infty$ and let q satisfy*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$(L^p(X))^* \cong L^q(X),$$

with duality pairing

$$\langle f, g \rangle = \int_X f(x) g(x) \, dx.$$

Thus:

$$\boxed{(L^p)^* = L^q \qquad (1 < p < \infty)}$$

C.6.3 Endpoint Cases

$p = 1$

$$(L^1(X))^* \cong L^\infty(X)$$

via the same pairing:

$$T_g(f) = \int_X f g \, dx.$$

$p = \infty$

$$(L^\infty(X))^* \supsetneq L^1(X).$$

The dual of L^∞ contains additional objects called *finitely additive measures* (the space $\mathrm{ba}(X)$). Hence:

$$(L^\infty)^* \neq L^1.$$

This is a fundamental difference between L^p theory and Radon measure theory.

C.6.4 Relation to Measures

Every function $f \in L^1(X)$ defines a Radon measure

$$d\mu = f(x)\, dx.$$

Thus:

$$L^1(X) \subset \mathcal{M}(X),$$

with strict inclusion in general.

However:

- $L^1(X)$ describes only absolutely continuous measures,
- $\mathcal{M}(X)$ also includes singular measures (e.g. Dirac masses),
- $C_0(X)^* = \mathcal{M}(X)$ is strictly larger than $(L^\infty)^*$.

C.6.5 Comparison of Dualities

Space	Dual	Notes
$C_0(X)$	$\mathcal{M}(X)$	Radon measures
$L^p(X),\ 1 < p < \infty$	$L^q(X)$	Reflexive
$L^1(X)$	$L^\infty(X)$	Non-reflexive
$L^\infty(X)$	$\mathrm{ba}(X)$	Not σ-additive

C.6.6 Conceptual Picture

$$\begin{array}{llll} C_0(X) & \xrightarrow{\text{dual}} & \mathcal{M}(X) & \text{(measures)} \\ L^p(X) & \xrightarrow{\text{dual}} & L^q(X) & \text{(functions)} \end{array}$$

- L^p duality is purely analytic.
- Measure duality is geometric/topological.
- Only for $p = 2$ do we get a Hilbert space structure.
- Measure duality captures singular objects; L^p duality does not.