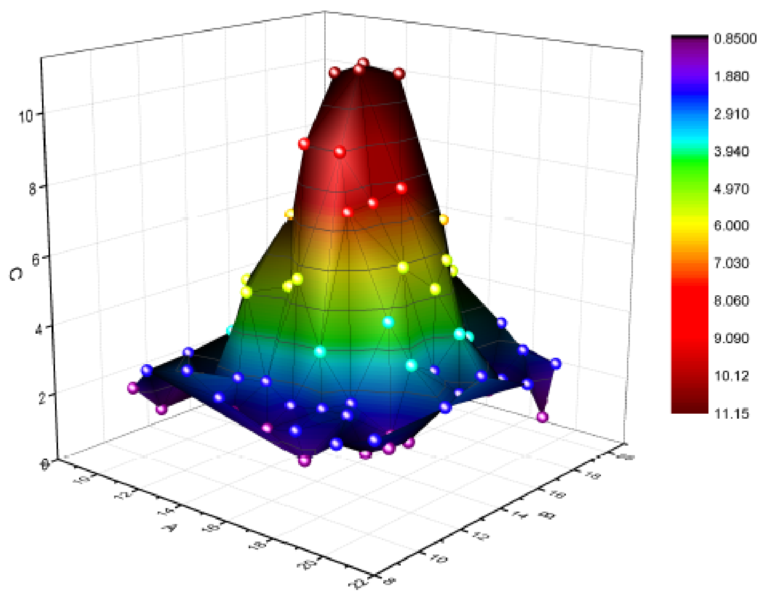
Themes of Statistics

Simple notions and simple proofs



Emilien Joly

Year 2020

Contents

Chapter 1

Preface

These notes, were essentially written during the first two years of my doctoral course at CIMAT, Mexico. As a student, I had the chance to have access to very well designed courses notes from my professors at the ENS Cachan and Université Paris Saclay which helped at lot in the learning process. This work is written in a way that it is as self-contained as I possibly achieved to, to quickly familiarize students with the beautiful notions around *empirical processes* and *Dudley entropy theory*.
These themes cannot be tackled without a quick tour by the classical convergence theorems in finite dimension spaces - where we speak about random vectors. This guided tour passes also rapidly through the simple 1D world as a excuse to look deeper into the important definitions in probability theory.
As a pedagogic material, this notebook pretends - I am aware of the gluttony for real life illustration asked by my students - to give enough instructive examples to get our hands on motivating application problems. [To continue]

Prerequisities: We assume known the following notions.

- Basic definitions of mathematical tools (sequences, integrals, limits, continuity, topology, limsup, liminf)
- σ-algebras, measurability, measures, probability measures, random variable, expected value, variance, independence, distributions.
- Classical theorems of integration (Monotone convergence, Dominated convergence, Fatou's Lemma,...)
- Classical distributions (Bernoulli, Binomial, Poisson, Exponential, Normal)

1.1 Notations and definitions

Vector space of finite dimension Let E be a vector space of finite dimension. As real vector spaces of same dimension are (linearly) equivalents, we will assume $E = \mathbb{R}^k$ for some $k \in \mathbb{N}$ fixed one and for all as it permits us to simplify our notations.

Sets of functions We denote by $\mathcal{C}_b(\mathbb{R}^k)$ the set of continuous and bounded functions $f : \mathbb{R}^k \mapsto \mathbb{R}$. For a measure μ on $\mathbb{R}^k$ and $p \geq 1$, we denote by $\mathbb{L}^p(\mathbb{R}^k, \mu)$ the set of measurable functions $f : \mathbb{R}^k \to \mathbb{R}$ such that $\int |f|^p d\mu < +\infty$. If μ is the Lebesgue measure on $\mathbb{R}^k$, the set $\mathbb{L}^p(\mathbb{R}^k, \mu)$ will be simply denoted by $\mathbb{L}^p(\mathbb{R}^k)$.

Notations o_P and O_P For a sequence of random vectors $(Z_n)_n$ and a sequence $(k_n)_n \in (\mathbb{R}_+)^{\mathbb{N}}$, we denote by

- $Z_n = O_P(k_n)$ if $\lim\limits_{T \to +\infty} \overline{\lim} \, \mathbb{P}\left(\|Z_n\| > Tk_n\right) = 0$,
- $Z_n = o_P(k_n)$ if for all $\varepsilon > 0$, $\lim\limits_{n \to +\infty} \mathbb{P}\left(\|Z_n\| > \varepsilon k_n\right) = 0$

Part I

Probability

Chapter 2

Convergence of random variables

The purpose of this chapter is to prepare the reader to enter in the field of empirical processes slowly by stating and proving the famous theorems as Law of Large Numbers (LLN) or Central Limit Theorem (CLT) which have made the popularity of Probability theory in the last century. A lot of this chapter is inspired by the excellent [15].

2.1 Modes of convergence

Definition 1. *A **random vector** is a random variable $X : \Omega \mapsto \mathbb{R}^k$ where we implicitly associated to Ω and $\mathbb{R}^k$ (with $k \in \mathbb{N}^*$) their respective Borelian σ-algebra. A sequence of random vectors will be usually denoted by $(X_n)_{n\in\mathbb{N}} \in (\mathbb{R}^k)^{\mathbb{N}}$.*

Definition 2. *Let $(X_n)_{n\in\mathbb{N}} \in (\mathbb{R}^k)^{\mathbb{N}}$ be a sequence of random vectors and X a random vector in $\mathbb{R}^k$. Their respective probability measures are denoted by μ_n and μ. Let d be a distance on $\mathbb{R}^k$ and $\|\cdot\|$ be the usual norm on $\mathbb{R}^k$. We say that,*

1. *$(X_n)_{n\in\mathbb{N}}$ converges in **probability** to X, denoted by $X_n \xrightarrow{\mathbb{P}} X$ if $\forall \epsilon > 0,\ \mathbb{P}\left(d(X_n, X) > \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0$.*

2. *$(X_n)_{n\in\mathbb{N}}$ converges in **distribution** or **weakly** to X, denoted by $X_n \xrightarrow{(d)} X$ or $X_n \xrightarrow{(w)} X$ if $\forall h \in \mathcal{C}_b(\mathbb{R}^k)$, $\int h d\mu_n \underset{n\to\infty}{\longrightarrow} \int h d\mu$.*

3. *$(X_n)_{n\in\mathbb{N}}$ converges in **almost surely** to X, denoted by $X_n \xrightarrow{a.s.} X$ if $\exists \Gamma \subset \Omega, \forall \omega \in \Gamma,\ X_n(\omega) \underset{n\to\infty}{\longrightarrow} X(\omega)$ and Γ^c is negligible.*

4. *$(X_n)_{n\in\mathbb{N}}$ converges **in** $\mathbb{L}^p$ to X, denoted by $X_n \xrightarrow{\mathbb{L}_p} X$ if $\forall n \in \mathbb{N}$, $\mathbb{E}\left[\|X_n\|^p\right] < +\infty$ and $\mathbb{E}\left[\|X_n - X\|^p\right] \underset{n\to\infty}{\longrightarrow} 0$.*

5. *$(X_n)_{n\in\mathbb{N}}$ converges in **total variation** to X, denoted $X_n \xrightarrow{TV} X$, if $\sup_B \left|\mathbb{P}\left(X_n \in B\right) - \mathbb{P}\left(X \in B\right)\right| \underset{n\to\infty}{\longrightarrow} 0$, where the supremum is taken over the set of Borelian measurable sets B.*

Remarks

- In 2., it is not required to have the random variables X_n and X to live in the same probability space whereas the other four type of convergence do require this fact.
- In 4., the triangular inequality implies $\mathbb{E}\left[\|X\|^p\right] < +\infty$.
- In the convergence in probability, since we are dealing with $\mathbb{R}^k$ (a vector space of finite dimension), all the distances are equivalent. This is to say, for any two distances d and d' on $\mathbb{R}^k$, there exists $c, C > 0$ such that, for every $x, y \in \mathbb{R}^k$
$$cd'(x, y) \leq d(x, y) \leq Cd'(x, y).$$
It implies that the notion of probability convergence that we consider is *not dependent* on the chosen distance. When not specified differently, we will always consider the euclidean distance.

The following Lemma simplifies the task of proving weak convergence and will be a key tool for the upcoming results.

Lemma 1 (Portmanteau)**.** *Let $(X_n)_{n\in\mathbb{N}}$ and X be random vectors. The following properties are equivalent:*

i) $X_n \xrightarrow{(d)} X$

ii) $\forall f$ *function Lipschitz and bounded,* $\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$.

iii) $\forall F$ *closed set,* $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$.

iv) $\forall G$ *open set,* $\liminf \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$.

v) $\forall A$ *Borelian s.t.* $\mathbb{P}(X \in \partial A) = 0$, $\mathbb{P}(X_n \in A) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \in A)$.

Proof. i) $\Longrightarrow$ ii) is obvious since Lipschitz bounded functions are in particular continuous and bounded.
ii) $\Longrightarrow$ iv) Let $f_k(x) = \min(kd(x, G^c), 1)$. This function is Lipschitz by the Lipschitzness of the distance. It is obviously bounded. Moreover, for every x, $f_k(x)$ converges increasingly to $\mathbb{1}_G(x)$. Hence,

$$\liminf_n \mathbb{P}(X_n \in G) \geq \liminf_n \mathbb{E}[f_k(X_n)] \overset{\text{by ii)}}{=} \mathbb{E}[f_k(X)] \underset{k\to\infty}{\longrightarrow} \mathbb{P}(X \in G)$$

where the last fact holds by monotone convergence.
iii)$\Leftrightarrow$ iv) is obvious by completion.
iii) + iv) $\Longrightarrow$ v) Take any Borelian set such that $\mathbb{P}(X \in \partial A) = 0$. Then, using iii) for the closed $\overline{A}$ and iv) for $\mathring{A}$, we get

$$\begin{array}{ccc} \limsup \mathbb{P}(X_n \in A) \leq \limsup \mathbb{P}\left(X_n \in \overline{A}\right) \leq & \mathbb{P}\left(X \in \overline{A}\right). \\ \rotatebox{90}{$\geq$} & \| \\ \liminf \mathbb{P}(X_n \in A) \geq \liminf \mathbb{P}\left(X_n \in \mathring{A}\right) \geq & \mathbb{P}\left(X \in \mathring{A}\right). \end{array}$$

This chain of inequalities finally imply that

$$\mathbb{P}(X_n \in A) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \in A).$$

v) $\Longrightarrow$ iii) Let F be a closed set of $\mathbb{R}^k$ and define for any $\beta > 0$,

$$F_\beta = \{x : d(x, F) \leq \beta\}.$$

The elements of the familly $(\partial F_\beta)_{\beta>0}$ are disjoint. Then

$$\sum_{\beta>0} \mathbb{P}(X \in \partial F_\beta) \leq \mathbb{P}\left(X \in \mathbb{R}^k\right) = 1.$$

The previous convergence has to be understood as the sumable (see Definition 13 and Proposition 23) then the sum has only finite number of non zero terms:

$$\{\beta > 0 : \mathbb{P}(X \in \partial F_\beta \neq 0) \text{ is a countable set.}$$

From that we can define a sequence $(\beta_k)_k$ such that $\beta_k \to 0$ and such that

$$\forall k \in \mathbb{N},\ \mathbb{P}(X \in \partial F_{\beta_k}) = 0.$$

Then

$$\limsup_{n\to+\infty} \mathbb{P}(X_n \in F) \leq \limsup_{n\to+\infty} \mathbb{P}(X_n \in F_{\beta_k}) = \lim_{n\to+\infty} \mathbb{P}(X_n \in F_{\beta_k}) \underset{\text{by v)}}{=} \mathbb{P}(X \in F_{\beta_k}).$$

We finish by taking the infimum in k.
iii) $\Longrightarrow$ i) Let $0 < f < 1$ be a continuous function. Using the classical (13.4) and Fatou Lemma, we get

$$\begin{aligned} \limsup \mathbb{E}[f(X_n)] &\leq \int_0^1 \limsup \mathbb{P}(f(X_n) \geq x)\, dx \\ &\underset{\text{by iii)}}{\leq} \int_0^1 \mathbb{P}(f(X) \geq x)\, dx = \mathbb{E}[f(X)]. \end{aligned}$$

We used that the set $\{f(X_n) \geq x\} = \{X_n \in f^{-1}([x, +\infty)\}$ where the set $f^{-1}([x, +\infty)$ is the inverse of a closed set and is then closed by continuity of f. Applying the same ideas for $1 - f$ gives the convergence

$$\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)].$$

Then the general case follows from this by using the transform $g := \frac{f-a}{b-a}$ for $a < f < b$. ☐

Many of the convergences of interest are robust under a continuous transformation. Precisely, we have the

Theorem 1 (Continuous transformation)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a continuous function. Then,*

- *If $X_n \xrightarrow{(d)} X$, then $g(X_n) \xrightarrow{(d)} g(X)$.*
- *If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.*
- *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.*

One could be interested in a result where g is only assumed to be continuous except on a specific set of points. The results are still true in this context if one assumes that this set of problematic points is not *seen* by the random variable X.

Proof. We prove in order:

- Let F be a closed set in $\mathbb{R}^m$. Then,

$$\begin{aligned}\limsup \mathbb{P}\left(g(X_n) \in F\right) &= \limsup \mathbb{P}\left(X_n \in g^{-1}(F)\right) \\ &\leq \mathbb{P}\left(X \in g^{-1}(F)\right) = \mathbb{P}\left(g(X) \in F\right).\end{aligned}$$

  which implies the weak convergence.

- Let $\varepsilon > 0$ and $\delta > 0$. We can decompose

$$\begin{aligned}\mathbb{P}\left(d(g(X_n), g(X)) > \varepsilon\right) \leq \mathbb{P}\left(d(g(X_n), g(X)) > \varepsilon \text{ and } d(X_n, X) \leq \delta\right) &\underset{\delta \to 0}{\longrightarrow} 0 \\ + \underbrace{\mathbb{P}\left(d(X_n, X) > \delta\right)}_{\underset{n\to\infty}{\longrightarrow} 0,\ \forall \delta > 0}&\end{aligned}$$

  This proves the convergence in probability.

- The almost sure convergence is obvious since it occurs on the same measurable set of probability 1.

□

2.1.1 Uniform integrability

Definition 3. *We say that a family $\mathcal{C}$ of random variables are **uniformly integrable at order p** (denoted U.I.) if $\forall \varepsilon > 0$, $\exists K \in [0; +\infty)$ such that*

$$\mathbb{E}\left[\|X\|^p \mathbb{1}_{\|X\|>K}\right] \leq \varepsilon, \forall X \in \mathcal{C}.$$

When $p = 1$ we omit to say "of order 1".

A U.I. family is bounded in $\mathbb{L}_p$ Take $\varepsilon = 1$ and we denote by K the constant defined in Definition 3. Then, for any element $X \in \mathcal{C}$, we have that

$$\mathbb{E}\left[\|X\|^p\right] \leq \mathbb{E}\left[\|X\|^p \mathbb{1}_{\|X\|^p>K}\right] + \mathbb{E}\left[\|X\|^p \mathbb{1}_{\|X\|^p\leq K}\right] \leq 1 + K.$$

Then a family that is uniformly integrable is, in particular, bounded in $\mathbb{L}_p$. Besides the following example allows us to see that the converse is not true.

Exercice 1. *Let $X_n = n\mathbb{1}_{[0,n^{-1})}$. Show that $\mathbb{E}[X_n] = 1$ and that $(X_n)_n$ is not U.I.*

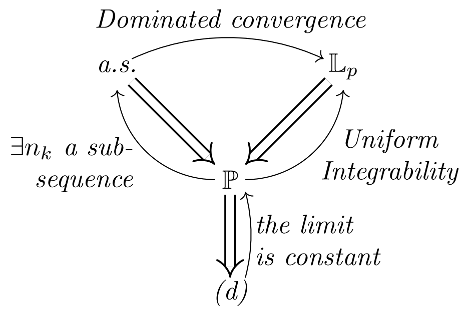
Sufficient conditions for U.I. There is two very simple sufficient conditions for uniform integrability that we state now.

Proposition 1. *If either*

- *The family $\mathcal{C}$ is bounded in $\mathbb{L}_{p'}$ for $p' > p$*
- *The family $\mathcal{C}$ is bounded by a random variable $Y \in \mathbb{L}_p$*

then $\mathcal{C}$ is uniformly integrable of order p.

Theorem 2 (Implication of convergences)**.** *We have the following implications for X_n and X random vectors in $\mathbb{R}^d$.*



The doubled arrows hold for direct consequences whereas the simple arrows hold with an extra assumption or in a weaker version has specified by the text aside. More specifically, we have the following results.

1. *Assume that* $X_n \xrightarrow{a.s.} X$ *and that there exists a random vector* Y *such that* $\|X_n\| \leq \|Y\|$ *for any* n *then* $X_n \xrightarrow{\mathbb{L}_p} X$.

2. *Assume that* $X_n \xrightarrow{\mathbb{P}} X$ *then there exists a sub-sequence* $(n_k)_k$ *such that* $X_{n_k} \xrightarrow{a.s.} X$.

3. *Assume that* $X_n \xrightarrow{\mathbb{P}} X$ *and that the familly* $(X_n)_n$ *is uniformly integrable at order* p *then* $X_n \xrightarrow{\mathbb{L}_p} X$.

4. *Assume that* $X_n \xrightarrow{(d)} c$ *where* c *is deterministic, then* $X_n \xrightarrow{\mathbb{P}} c$.

Proof. a.s. $\implies \mathbb{P}$: We assume that $X_n \xrightarrow{a.s.} X$.

$$\begin{aligned} 0 &= \mathbb{P}\left(\exists \text{ a sub-sequence } n_k \text{ s.t. } \forall k, |X_{n_k} - X| > \varepsilon\right) \\ &= \mathbb{P}\left(\limsup \{|X_n - X| > \varepsilon\}\right) \qquad \text{(seen as events)} \\ &\geq \limsup \ \mathbb{P}\left(|X_n - X| > \varepsilon\right) \qquad \text{(Fatou for events)} \end{aligned}$$

and then $\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0$ for any $\varepsilon > 0$.
$\mathbb{P} \implies$ (d): Let f be a λ-Lipschitz function bounded by a constant K, then

$$\begin{aligned} |\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]| &\leq \mathbb{E}\left[|f(X_n) - f(X)| \mathbb{1}_{|X_n - X| \leq \varepsilon}\right] + 2K\mathbb{P}\left(|X_n - X| > \varepsilon\right) \\ &\leq \lambda\varepsilon + 2K\mathbb{P}\left(|X_n - X| > \varepsilon\right) \end{aligned}$$

The convergence in probability allows us to choose n large enough to get $\mathbb{P}\left(|X_n - X| > \varepsilon\right) \leq \varepsilon$. Then $|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]| \leq (\lambda + 2K)\varepsilon$ which shows that $\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]$. We conclude using Lemma 1 to get the weak convergence.
$\mathbb{L}_p \implies \mathbb{P}$: By the Markov's inequality,

$$\mathbb{P}\left(\|X_n - X\| > \varepsilon\right) \leq \frac{\mathbb{E}\left[\|X_n - X\|^p\right]}{\varepsilon^p} \underset{n\to\infty}{\longrightarrow} 0$$

1. a.s. $\to \mathbb{L}_p$ is the direct consequence of the dominated convergence theorem. Indeed, by the bounded condition, X is in $\mathbb{L}_p$ and $\|X\| \leq \|Y\|$. Then we get

$$\|X_n - X\| \leq \|Y\| + \|X\| \leq 2\|Y\|$$

which is in $\mathbb{L}_p$. Using, the dominated convergence theorem for the sequence $(\|X_n - X\|^p)_n$ finally gives the result.
2. $\mathbb{P} \to$ a.s. This fact results from an interesting result in itself that we postpone to Lemma 36.
3. $\mathbb{P} \to \mathbb{L}_p$ For simplicity, we show the result for $p = 1$ and $X_n \in \mathbb{R}$ since the generalization to any p and $X_n \in \mathbb{R}^k$ is straightforward. Let $\phi_K : \mathbb{R} \to [-K, K]$ such that

$$\phi_K := \begin{cases} K & \text{if } x > K \\ x & \text{if } |x| \leq K \\ -K & \text{if } x < -K \end{cases} .$$

Let $\varepsilon > 0$. Since the family $(X_n)_n$ is U.I., there exists $K > 0$ such that

$$\mathbb{E}\left[|\phi_K(X_n) - X_n|\right] < \frac{\varepsilon}{3} \quad \forall n \geq 0,$$

and

$$\mathbb{E}\left[|\phi_K(X) - X|\right] < \frac{\varepsilon}{3}.$$

By construction ϕ_k is 1-Lipschitz i.e. $\forall x, y,\ |\phi_K(x) - \phi_K(y)| \le |x-y|$ then by the continuous transformation

$$\phi_K(X_n) \xrightarrow{\mathbb{P}} \phi_K(X).$$

We can use the dominated convergence theorem (see Lemma 26) since $\phi_K(X_n)$ and $\phi_K(X)$ are bounded (and then integrable) to see that there exists n_0 such that $\forall n \ge n_0$,

$$\mathbb{E}\left[|\phi_K(X_n) - \phi_K(X)|\right] < \frac{\varepsilon}{3}.$$

Summing up, we get

$$\mathbb{E}\left[|X_n - X|\right] \le \mathbb{E}\left[|X_n - \phi_K(X_n)|\right] + \mathbb{E}\left[|\phi_K(X_n) - \phi_K(X)|\right] + \mathbb{E}\left[|\phi_K(X) - X|\right] < \varepsilon.$$

Then $X_n \xrightarrow{\mathbb{L}_p} X$.
4. (d) $\to \mathbb{P}$ Let $B(c,\varepsilon)$ be the open ball of radius ε centered at c. Then $\mathbb{P}\left(d(X_n, c) \ge \varepsilon\right) = \mathbb{P}\left(X_n \in B(c,\varepsilon)^c\right)$, but

$$\limsup \mathbb{P}\left(X_n \in B(c,\varepsilon)^c\right) \le \mathbb{P}\left(c \in B(c,\varepsilon)^c\right) = 0,$$

by the lemma Portmanteau. Hence, $\mathbb{P}\left(d(X_n, c) \ge \varepsilon\right) \to 0$ and $X_n \xrightarrow{\mathbb{P}} c$. □

Two exercises about probability convergence

Exercice 2. *Define the sequence of random variables on the probability triplet* $((0,1], \mathcal{B}((0,1]), Leb)$,

$$\begin{aligned}
&Y_1 = \mathbb{1}_{(0,1]} \\
&Y_2 = \mathbb{1}_{(0,1/2]}, Y_3 = \mathbb{1}_{(1/2,1]} \\
&Y_4 = \mathbb{1}_{(0,1/4]}, Y_5 = \mathbb{1}_{(1/4,1/2]}, Y_6 = \mathbb{1}_{(1/2,3/4]}, Y_7 = \mathbb{1}_{(3/4,1]} \\
&\dots
\end{aligned}$$

Show that this sequence is such that $Y_n \xrightarrow{\mathbb{P}} 0$ *but has no almost sure limit. We list its basic properties in the following proposition.*

Exercice 3. *Let* X_n *be a sequence of random variables that converges in probability towards a random variable* X*. Assume that* $\forall n \in \mathbb{N}, X_n \le X_{n+1}$*. Show that* $X_n \xrightarrow{a.s.} X$*. Hint: Use 2. of Theorem 2.*

Comments In fact the convergence $\mathbb{L}_p$ implies a little more than the convergence in probability. It also implies the uniform integrability as pledged in Exercice 4.

Exercice 4 ($\mathbb{L}_p \implies$ U.I.)**.** *Assume that* $X_n \xrightarrow{\mathbb{L}_p} X$*. We show in that exercise that* $(X_n)_n$ *is uniformly integrable of order* p*.*

1. *Let* $\varepsilon > 0$*. Show that there exists* $N \in \mathbb{N}$ *such that* $\forall n \ge N$, $\mathbb{E}\left[\|X_n - X\|^p\right] \le \varepsilon/2^p$*.*

2. *Apply Proposition 24 to show that we can choose* $\delta > 0$ *such that for any* $E \in \mathcal{B}$ *such that* $\mathbb{P}(E) < \delta$*, we have*

$$\mathbb{E}\left[\|X_n\|^p \mathbb{1}_E\right] \le \varepsilon/2^{p-1},\ \forall n \le N \qquad \textit{and} \qquad \mathbb{E}\left[\|X\|^p \mathbb{1}_E\right] \le \varepsilon/2^p$$

3. *Taking* K *such that* $K^{-1}\sup_n \mathbb{E}\left[\|X_n\|^p\right] \le \delta$*, show that* $(X_n)_n$ *is U.I. using that,*

$$\mathbb{E}\left[\|X_n\|^p 1\|X_n\| > K\right] \le 2^{p-1}\mathbb{E}\left[\|X\|\mathbb{1}_{\|X_n\|>K}\right] + 2^{p-1}\mathbb{E}\left[\|X_n - X\|^p\right],$$

   (We may use Lemma 20) for $n > N$ *and question 2. for* $n \le N$*.*

2.1.2 Simultaneous convergence

In this section, we deal with the simultaneous convergence of two random variables X_n and Y_n when it is known that they marginally converge to two random variables X and Y. Combining their convergence is not that direct, especially for weak convergence. In the following, the famous Slutsky Lemma is also presented as an optimal result in this direction.

Convergence almost sure Almost nothing is needed to say here. Indeed, considering the intersection of the two measurable sets on which $X_n(\omega) \to X(\omega)$ and $Y_n(\omega) \to Y(\omega)$ results another set of probability one where simultaneously the two convergences occur. Simultaneous convergence being equivalent to convergence for the sequence of couples in product spaces gives the result. We keep that in mind under the short,

$$X_n \xrightarrow{a.s.} X \text{ and } Y_n \xrightarrow{a.s.} Y \Leftrightarrow (X_n, Y_n) \xrightarrow{a.s.} (X, Y)$$

Convergence in probability By the fact that for x_1, y_1, x_2, y_2, we have (for the euclidean distance)

$$d((x_1, y_1), (x_2, y_2)) \leq d(x_1, x_2) + d(y_1, y_2),$$

and for example,

$$d(x_1, x_2) \leq d((x_1, y_1), (x_2, y_2))$$

then, the probability convergence transmits directly in product spaces. More precisely,

$$X_n \xrightarrow{\mathbb{P}} X \text{ and } Y_n \xrightarrow{\mathbb{P}} Y \Leftrightarrow (X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y)$$

Slutsky Lemma

Proposition 2. *Let $(X_n)_n$ and $(Y_n)_n$ be two sequences of random vectors. Assume that $X_n \xrightarrow{(d)} X$ and $d(X_n, Y_n) \xrightarrow{\mathbb{P}} 0$, then $Y_n \xrightarrow{(d)} X$.*

Proof. Let f be a 1-Lipschitz function taking values in $[0, 1]$. Note that imposing f to take values in $[0, 1]$ is not restrictive since one can always renormalize and translate a bounded function. Then,

$$\begin{aligned} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(Y_n)]| &\leq \mathbb{E}\left[d(X_n, Y_n)\mathbb{1}_{d(X_n,Y_n)\leq\varepsilon}\right] + 2\mathbb{P}(d(X_n, Y_n) > \varepsilon) \\ &\leq \varepsilon + 2\underbrace{\mathbb{P}(d(X_n, Y_n) > \varepsilon)}_{\underset{n\to\infty}{\longrightarrow} 0}. \end{aligned}$$

Then, $\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$ and the weak convergence is proved. □

The so-called Slutsky Lemma is very useful in many areas of statistics as a powerful tool to combine the convergence of two or more sequence of random variables to finally get the weak convergence of a possibly complex expression.

Lemma 2 (Slutsky)**.** *Assume that $X_n \xrightarrow{(d)} X$ and $Y_n \xrightarrow{\mathbb{P}} c$ where c is a constant of $\mathbb{R}^k$. Then, $(X_n, Y_n) \xrightarrow{(d)} (X, c)$ and in particular we have*

- $X_n + Y_n \xrightarrow{(d)} X + c$.
- $Y_n X_n \xrightarrow{(d)} cX$.
- $Y_n^{-1} X_n \xrightarrow{(d)} c^{-1} X$ *when* $c \neq 0$.

Proof. We use the previous proposition with $(X_n, c) \xrightarrow{(d)} (X, c)$ and $d((X_n, c), (X_n, Y_n)) \leq d(Y_n, c) \underset{n\to\infty}{\longrightarrow} 0$ where we used indistinctly d for the distance in $\mathbb{R}^k$ and $\mathbb{R}^{2k}$. □

Exercice 5. *Prove that $X_n \xrightarrow{(d)} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$ is not sufficient (in general) to have $(X_n, Y_n) \xrightarrow{(d)} (X, Y)$. (Hint: Consider $X_n = Y_n = Y$ and $X \sim Y$ drawn independently.)*

The particular case follows from the continuous transformation of the weak convergence.

Example of application of Slutsky Lemma If one takes $X_1, \dots, X_n$ a collection of i.i.d. random vectors such that $\mathbb{E}[X_1] = 0$ and $\mathbb{E}\left[X_1^2\right] < +\infty$. One can compute the two classical estimators,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

By the weak law of large numbers, $\overline{X}_n \xrightarrow{\mathbb{P}} 0$ and

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right) \xrightarrow{\mathbb{P}} \mathbb{E}\left[X_1^2\right] - (\mathbb{E}\left[X_1\right])^2 = \mathrm{Var}\left(X_1\right)$$

where we used Theorem 1 for the function $g(x,y) = x - y^2$. The central limit theorem also gives that $\sqrt{n}\,\overline{X}_n \xrightarrow{(d)} \mathcal{N}(0, \mathrm{Var}\left(X_1\right))$ which, combined with Slutsky's Lemma, implies

$$\sqrt{n}\,\frac{\overline{X}_n}{S_n^2} \xrightarrow{(d)} \mathcal{N}(0,1).$$

This last property allows to design confidence intervals for the mean $\mathbb{E}\left[X_1\right]$ of a sample of unknown common variance.

2.2 Exercices

Exercice 6. *Let $(X_n)_{n\geq 0}$ a sequence of real random variables.*

1. *Show that the convergence in distribution of $(X_n)_{n\geq 1}$ is NOT equivalent to " For any continuous function of compact support f, the sequence $(\mathbb{E}(f(X_n)))_{n\geq 1}$ converge."*

2. *Show that the convergence in distribution of $(X_n)_{n\geq 1}$ is equivalent to "For any continuous function of compact support f, the sequence $\mathbb{E}(f(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X_0))$."*

3. *We assume that $X_n \xrightarrow[n\to\infty]{L^1} X_0$.*

   (a) Show that for any fixed $\epsilon > 0$, there exists $\delta > 0$ such that $\mathbb{E}(\|X_n\|\mathbf{1}_{X_n\in F}) < \epsilon$ for all $n \geq 0$ and any $F \in \mathcal{B}(\mathbb{R})$ such that $\mathbb{P}(F) \leq \delta$.

   (b) Deduce that if $X_n \xrightarrow[n\to\infty]{L^1} X_0$, then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X_0$ y $(X_n)_{n\geq 0}$ is uniformly integrable.

Exercice 7. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables.*

1. *Assume that $(X_n)_{n\geq 1}$ converges in distribution to a standard gaussian random variable N. Is there convergence of $\mathbb{E}(|X_n|^p)$ towards $\mathbb{E}(|N|^p)$ for any $p \geq 1$?*

2. *Show the converse: If the sequence $\mathbb{E}(|X_n|^p)$ converges to $\mathbb{E}(|N|^p)$ for all $p \geq 1$, then $(X_n)_{n\geq 1}$ converges in distribution to the standard gaussian variable N.*

Exercice 8. *Let $(X_n)_{n\geq 1}$ be a sequence of real random variables with support included in $\mathbb{Z}$.*

1. *We assume that $(X_n)_{n\geq 1}$ converges in distribution towards X. What is the support of X? Show that for any $x \in \mathbb{Z}$,*

$$\mathbb{P}(X_n = x) \xrightarrow[n\to\infty]{} \mathbb{P}(X = x).$$

2. *Assume that X is a real random variable and that for all $x \in \mathbb{Z}$,*

$$\mathbb{P}(X_n = x) \xrightarrow[n\to\infty]{} \mathbb{P}(X = x).$$

   What should verify X so that X_n converges to X ?

Exercice 9. *Let $(X_n)_{n\geq 1}$ be a sequence of binomial random variables of parameters $(n, 1/n)$. Let $(Y_n)_{n\geq 1}$ be a sequence of random variables such that for any $x \leq \sqrt{n}$, conditionally to $X_n = x$, we have that $Y_n = x$ and otherwise, conditionally to $X_n = x$, we have that Y_n is a binomial random variable of parameters $(x!, \frac{1}{\pi})$. Show that $(Y_n)_{n\geq 0}$ converges in distribution and describe the limit.*

Exercice 10. *Let X be a random variable of support included in $\mathbb{Z}$ and with distribution*

$$\mathbb{P}(X = n) = \frac{C}{2n^2 \log n},$$

for all $n \in \mathbb{Z}$.

1. *Show that X has no moment of order 1.*
2. *Calculate the characteristic function ϕ_X of X.*
3. *Show that ϕ_X is differentiable on $\mathbb{R}$.*

Exercice 11. *Let Z be a random variable with uniform distribution on $[-1, 1]$.*

1. *Compute the characteristic function of Z.*
2. *Show that there is no i.i.d. random variables X, Y such that $X - Y \sim Z$.*

Chapter 3

Distribution function

For a random vector $X = (X_1, \dots, X_k)$, the function $F_X : \mathbb{R}^k \to [0,1]$ and given by

$$F_X(x_1, \dots, x_k) = \mathbb{P}(X_1 \le x_1, \dots, X_k \le x_k)$$

is called the **distribution function** of the random vector X. In the real case, it is obvious to see that the distribution function is no-decreasing. The vectorial case is a little different in the notion of monotonicity of F_X. We say that a function f is **2-increasing** if for any two coordinate i and j in $\{1, \dots, k\}$, we have $\forall x \le y$ and $\forall u \le v$,

$$\Delta^{(i)}_{x,y} \Delta^{(j)}_{u,v} f \ge 0,$$

where $\Delta^{(i)}_{a,b} = (f^{(i)}(\cdot, b) - f^{(i)}(\cdot, a))/(b-a)$ and $f^{(i)}(\cdot, x)$ holds for the function

$$(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k) \mapsto f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_k).$$

Proposition 3. *We have the following. For two vectors x, y of $\mathbb{R}^k$, we denote by $x \le y$ if each coordinate of x is smaller than each coordinate of y.*

a) F_X is a 2-increasing function.

b) Denoting by $x \to +\infty^k$ the fact that each coordinate of x tend to $+\infty$ and by $x \to -\infty^{\cup k}$ the fact that at least one of the coordinates converges to $-\infty$, we have that

$$\lim_{x \to +\infty^k} F_X(x) = 1 \qquad \text{and} \qquad \lim_{x \to -\infty^{\cup k}} F_X(x) = 0.$$

c) F_X is right-continuous.

Proof. Obvious. □

Remark 1. *The notion of right continuity is to be understood in its weak version. It is formally defined as*

'For any sequence $(x_n)_n \in (\mathbb{R}^k)^{\mathbb{N}}$ decreasing (coordinate by coordinate) to x, $F_X(x_n) \underset{n \to +\infty}{\longrightarrow} F_X(x)$'

A natural question is to ask whether or not those are the maximal properties that a distribution function have in full generality. We can answer by the affirmative thanks to the following section.

3.1 Existence of random variables of given distribution function

In this section, we will use the important Carathéodory extension theorem. See Theorem 16

Proposition 4. *Let $F : \mathbb{R}^k \to [0,1]$ which satisfies a),b) and c) of Proposition 3 then there exists a random vector $X \in \mathbb{R}^k$ such that $F_X = F$.*

Proof. We treat the case $k = 2$ since the general case is a direct generalization of this case. Assume given the function $F : \mathbb{R}^2 \to [0,1]$ and let Σ_0 be the algebra (in the sense of Definition 15.1.1) of all the sets which are Cartesian product of sets of the form

$$(a, b], \ (-\infty, b], \ (a, +\infty), \ \mathbb{R}, \ \emptyset \qquad \text{where } a, b \in \mathbb{R}.$$

One can directly construct a countably additive map $\mu_0 : \Sigma_0 \to [0,1]$ corresponding to the natural meaning of a distribution function. For example for the set $A = (a,b] \times (c,d]$ (where $a \le b$ and $c \le d$ with $a,b,c,d \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$) corresponding to a event of the form

$$\{a < X_1 \le b \ \&\ c < X_2 \le d\},$$

one would associate the value $\mu_0(A) := F(b,c) - F(b,c) - (F(a,d) - F(a,c))$. The first property of Proposition 3 implies that $\mu_0(A)$ is always a positive quantity. Also note that, in order to be consistent, we need the conditions $F(-\infty,\cdot) = F(\cdot,-\infty) = 0$ that are given by the second point of Proposition 3. The countably additive property of μ_0 follows easily from the right-continuous property of F. Hence Carathéodory theorem allows us to extend μ_0 to the σ-algebra generated by Σ_0 which is the Borelian sets. Hence, one have constructed a measure on $\mathbb{R}^2$ (and hence a corresponding random variable X) such that μ_X has distribution function F. □

In the following result, we state and prove a Lemma that is at the basis of the characterization of the convergence in distribution by the distribution functions.

Lemma 3 (Helly)**.** *Let $(F_n)_n$ be a sequence of distribution functions on $\mathbb{R}^k$. Then, there exists a non decreasing right-continuous function F such that $0 \le F \le 1$ and a sub-sequence $(n_i)_i$ such that*

$$\lim_{i\to\infty} F_{n_i}(x) = F(x) \quad \textit{for each point } x \textit{ of continuity of } F.$$

Be careful Lemma 3 is not sufficient to ensure that the resulting object F is a distribution function. Indeed, it is completely possible to be facing a case where

$$\lim_{x\to-\infty^k} F(x) \neq 0 \text{ or } \lim_{x\to\infty^k} F(x) \neq 1.$$

This comes from the fact that $\mathcal{P}(\mathbb{R}^k)$ is not compact in general. One can see that by considering the sequence $(\mu_n)_n$ such that $\mu_n = \delta_{(n,\ldots,n)}$ which has no sub-sequence that converges to a probability measure. Besides, the interested reader may be pleased to know that Riesz representation theorem makes of $\mathcal{P}(\overline{\mathbb{R}^k})$ (embedded with the weak topology) a compact metric space.

The following definition makes clear the suitable assumption to make to avoid dealing with the non-closed case of Helly's lemma.

Definition 4 (tension of measures)**.** *A sequence $(\mu_n)_n$ in $\mathcal{P}(\mathbb{R}^k)$ is said to be **tight** if*

$$\forall \varepsilon > 0, \exists K > 0 \ \textit{s.t. for all } n, \mu_n([-K,K]^k) \ge 1 - \varepsilon$$

Note that for the measures of a sequence of random vectors $(X_n)_n$, the previous definition is equivalent to

$$\lim_{x\to+\infty} \sup_n \mathbb{P}\left(\|X_n\| \ge x\right) = 0.$$

Exercice 12. *Show that the last assertion is true.*

We have the final

Lemma 4. *Let $(F_n)_n$ be a sequence of distribution functions on $\mathbb{R}^k$ such that*

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \textit{for each point } x \textit{ of continuity of } F.$$

Assume furthermore that $(F_n)_n$ is tight. Then, F is a distribution function on $\mathbb{R}^k$.

Proof. Since for all n, $F_n(K) \ge \mu_n([-K,K]^k) \ge 1-\varepsilon$, it holds that

$$\lim_{x\to+\infty^k} F(x) = 1$$

For any $x = (x_1,\ldots,x_k) \in \mathbb{R}^k$, $F_n(-K-1, x_2, \ldots, x_k) = \mu_n((-\infty,-K-1] \times (-\infty,x_2] \times \cdots \times (-\infty,x_k])$. But since the two sets $(-\infty,-K-1] \times (-\infty,x_2] \times \cdots \times (-\infty,x_k]$ and $[-K,K]^k$ are disjoints, we have

$$\mu_n((-\infty,-K-1] \times (-\infty,x_2] \times \cdots \times (-\infty,x_k]) \le 1 - \mu_n([-K,K]^k) \le \varepsilon,$$

and then

$$\lim_{x\to-\infty^{\cup k}} F_X(x) = 0.$$

□

The counter example fails to verify the tension condition as state in the following exercise.

Exercice 13. *Show that $\mu_n = \delta_{(n,\dots,n)}$ is not tight.*

Proof of Helly's Lemma. We have the inclusion of the countable set $\mathbb{Q}^k \subset \mathbb{R}^k$. Let $q_1, q_2, \dots$ be a enumeration of the elements of $\mathbb{Q}^k$. The sequence $(F_n(q_1))_n$ is a bounded sequence of (in $[0,1]$) reals. Then, by compactness, one can extract a sub-sequence such that $F_{n(1,j)}(q_1) \longrightarrow H(q_1)$ where the notations $n(1,j)$ and $H(q_1)$ hold respectively for the extractor sequence and for the limit. Now, the sequence $(F_{n(1,j)}(q_2))_j$ is also a sequence in $[0,1]$ and another extraction $n(2,j) \subset n(1,j)$ gives that $F_{n(2,j)}(q_2) \longrightarrow H(q_2)$. Hence one can construct a sequence of extraction such that

$$\forall i,\ F_{n(i,j)}(q_i) \underset{j\to\infty}{\longrightarrow} H(q_i).$$

We finally have that $\forall q \in \mathbb{Q}^k$, $H(q) = \lim_{i\to+\infty} F_{n(i,i)}(q)$. It is obvious to see that $0 \le H \le 1$ and that H is a 2-increasing function on $\mathbb{Q}^k$. We define, $\forall x \in \mathbb{R}^k$, $F(x) := \underset{q\downarrow x}{H}(q)$ it always exists since it is the limit of a decreasing sequence. It may not be clear that the function F is well defined. Let $(q_n)_n$ and $(q'_n)_n$ be two sequences such that $q_n \downarrow x$ and $q'_n \downarrow x$ and let $F(x)$ be the limit defined by $(q_n)_n$ and $F'(x)$ be the limit defined by $(q'_n)_n$. By the fact that $q_n \to x$, one can extract a sub-sequence q_{n_i} such that $\forall i$, $q_{n_i} \le q'_i$. Now, taking the limit in i, of $H(q_{n_i}) \le H(q'_i)$ gives $F(x) \le F'(x)$. But symmetrically, $F'(x) \le F(x)$ and the function F is well-defined. By construction, we have that F is right-continuous and,

$$F_{n(i,i)}(x) \longrightarrow F(x) \quad \text{for every point of continuity of } F.$$

□

When the limiting function F is continuous, we have a stronger result.

Proposition 5 (Glivenko-Cantelli)**.** *Let $(X_n)_n$ be a sequence of random variables in $\mathbb{R}$ of distribution function $(F_n)_n$. Assume that $X_n \overset{(d)}{\longrightarrow} X$ where we denote by F the distribution function of X. Assume that F is continuous on $\mathbb{R}$, then*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. Let $m \in \mathbb{N}^*$ and let $-\infty = x_0 < x_1 < \cdots < x_m = +\infty$ such that $F(x_i) = i/m$. This is possible since F is continuous. (The x_i may not be unique.) Then, for any $x \in [x_{i-1}, x_i]$,

$$F_n(x) - F(x) \le F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{m}.$$

In the same way, we have that $F_n(x) - F(x) \ge F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}$. From those two facts, we have that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \sup_{0\le i\le m} |F_n(x_i) - F(x_i)| + \frac{1}{m}.$$

Now, let $\varepsilon > 0$ and fix $m \le 2/\varepsilon$ such that $1/m \le \varepsilon/2$. Remark that the supremum is taken over a finite family of random variables so the classical law of large numbers (Proposition 10) can be applied $m+1$ times to get that for n large enough,

$$\sup_{0\le i\le m} |F_n(x_i) - F(x_i)| \le \frac{\varepsilon}{2}.$$

This concludes the proof. □

Chapter 4

Levy theorem

Levy's theorem is one of the building blocks of the study of characteristic functions. It characterizes the convergence in law of random variables through the convergence of their Fourier transforms. It is one of the simplest way to prove the CLT for random vectors. Before going through the theorem itself, one need to develop a few tools in the area of functional analysis and Fourier transform in L^p.

4.1 Characteristic function

For a random variable X of measure μ, the function defined for any $t \in \mathbb{R}^k$,

$$\phi_X(t) = \mathbb{E}\left[\exp(it \cdot X)\right]$$

is called **characteristic function**. This notion is deeply linked with functional analysis. Indeed, the **Fourier transform of a measure** is defined as

$$\mathcal{F}\mu(\xi) = \int_{\mathbb{R}^k} \exp(-ix \cdot \xi) d\mu(x).$$

so that we have $\phi_X(t) = \mathcal{F}\mu(-t)$. From this fact, all the properties that are possible to show on the Fourier transform can be settled for characteristics functions and vice versa. Some authors like to presents ad hoc proofs on characteristic functions. We choose to write things in a way that it is close in notation and spirits to the functional analysis literature.

4.1.1 Basic properties of the characteristic function

Proposition 6. *Let X be a random vector and let ϕ_X be its characteristic function. We have the following facts.*

1. $\phi_X(0) = 1$.
2. *For all $t \in \mathbb{R}^k$, $|\phi_X(t)| \leq 1$.*
3. *On $\mathbb{R}^k$, the function $t \mapsto \phi_X(t)$ is continuous.*
4. *For any $a \in \mathbb{R}$ and $b \in \mathbb{R}^k$, $\phi_{aX+b}(t) = e^{ib \cdot t}\phi_X(at)$.*
5. *If for $n \in \mathbb{N}$, $\mathbb{E}\left[\|X\|^n\right] < \infty$, we have*

$$\partial_j^{(n)}\phi_X(t) = \mathbb{E}\left[(iX_j)^n e^{it\cdot X}\right]$$
$$\text{and} \quad \partial_j^{(n)}\phi_X(0) = i^n \mathbb{E}\left[(X_j)^n\right]$$

Proof. All the statement are simple use of classical results in integration as dominated convergence theorems. □

It is important to know that most of the classical distribution have explicit formulas for the characteristic function.

Example 1. *The caracteristic function of $\mathcal{N}(\mu, \sigma^2)$ is*

$$\forall t \in \mathbb{R}, \quad \phi_{\mu,\sigma^2}(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right).$$

Proof. A random variable $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ can be written $X = \mu + \sigma Z$ where $Z \sim \mathcal{N}(0,1)$. So, $\phi_{\mu,\sigma^2}(t) = e^{it\mu}\phi(\sigma t)$ where ϕ is the characteristic function of Z. It is sufficient to prove $\phi(t) = e^{-t^2/2}$. Since the density function $f_{0,1}$ of $\mathcal{N}(0,1)$ is symmetric, we have that $\forall t \in \mathbb{R}$, $\phi(t) = \phi(-t)$ hence,

$$\phi(t) = \frac{\phi(t)+\phi(-t)}{2} = \int_{\mathbb{R}} \frac{e^{itz}+e^{-itz}}{2} f_{0,1}(z)dz = \int_{\mathbb{R}} \cos(tz)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

and then $\phi(t)$ is real. By the theorem of derivation under the integral and integration by parts,

$$\phi'(t) = \int_{\mathbb{R}} \sin(tz)\frac{-z}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz = -\int_{\mathbb{R}} t\cos(tz)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz = -t\phi(t).$$

This simple linear equation takes as solutions the functions $\phi(t) = e^{-t^2/2} + C$, but $\phi(0) = 1$ then $C = 0$. Finally, the only possibility is $\phi(t) = e^{-t^2/2}$. □

4.2 Fourier analysis

4.2.1 Convolution of measures

For μ probability measure (see [9] for more general measures) and f a function integrable with respect to μ, we define the **convolution of a function by a measure** $f \star \mu$ by

$$f \star \mu : x \mapsto \int_{\mathbb{R}^k} f(x-y)d\mu(y).$$

Also, the **convolution between two measures** μ and ν is given by

$$\forall A \text{ measurable}, \quad \mu \star \nu(A) = \int_{\mathbb{R}^k \times \mathbb{R}^k} \mathbb{1}_{x+y\in A} d\mu(x)d\nu(y)$$

where $\mathcal{A}$ and $\mathcal{B}$ are the respective σ-algebras of μ and ν. It will be checked in the appendix that $\mu \star \nu$ is indeed a probability measure on $\mathbb{R}^k$ in Fact 1. It is shown in appendix the habitual:

Proposition 7. *The Fourier transform satisfies the following basic properties. For μ and ν two probability measures,*

- $\|\mathcal{F}\mu\|_\infty \leq 1$.
- $\mathcal{F}(\mu \star \nu) = (\mathcal{F}\mu) \times (\mathcal{F}\nu)$.

The convolution of measures is very convenient to compute the distribution of sums of independent random variables.

Proposition 8. *Let $X \sim \mu$ and $Y \sim \nu$ be two independent random variables and let $Z = X + Y$. Then*

i) Z has the probability law given by $\mu \star \nu$.

ii) If X has a continuous bounded density f, then Z has a continuous density given by $f \star \nu$.

The second fact can be useful when one wants to smooth some distribution Y by a small X in order to get a random variable Z that has a density.

Proof. Point $i)$ can be seen on all borelians of the form $(-\infty, a]$, for example. Point $ii)$ can be seen using that $\forall h$ lipschitz,

$$\mathbb{E}\left[h(Z)\right] = \mathbb{E}\left[h(X+Y)\right] = \iint h(x+y)f(x)dxd\nu(y) = \int h(z)\left(\int f(z-y)d\nu(y)\right)dz.$$

□

4.2.2 Inversion formula

Parseval Identity Let X and Y be two random variables taking values in $\mathbb{R}^k$ of respective measures μ and ν. Finally, we denote by ϕ_μ the characteristic function of X and by ϕ_ν the characteristic function of Y. We get that, for any $t \in \mathbb{R}^k$

$$\exp(-i\xi \cdot t)\phi_\mu(\xi) = \int_{\mathbb{R}^k} \exp(i\xi \cdot (x-t))d\mu(x).$$

Under the condition that $\phi_\mu \in L_1(\mathbb{R}^k, \nu)$ (integrable with respect to ν), integrating both sides with respect to ν and using Fubini's theorem give that

$$\int_{\mathbb{R}^k} \exp(-i\xi \cdot t)\phi_\mu(\xi)d\nu(\xi) = \int_{\mathbb{R}^k} \phi_\nu(x-t)d\mu(x). \tag{4.1}$$

This equation is called **Parseval inequality**. It has to be understood as a continuous version of the Perseval inequality for periodic functions. As for Fourier series, it is a inversion formula that permits to link the norms of the transform of a function (here the characteristic function) and of the function itself.

Special case When one specify the previous identity where one takes ν to be a normal probability measure, centered and of variance $\sigma^{-2}I$, the previous identity takes the form

$$\frac{\sigma^k}{(2\pi)^{k/2}} \int_{\mathbb{R}^k} \exp(-i\xi \cdot t)\phi_\mu(\xi)\exp(-\frac{1}{2}\sigma^2\xi^2)d\xi = \int_{\mathbb{R}^k} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right)d\mu(x)$$

where the square of a vector has to be understood as the square of its norm.

Inversion Formula We are now ready to give the complete proof of the inversion formula.

Theorem 3. *Let μ be a borelian measure of probability on $\mathbb{R}^k$ let X be a random variable of measure μ. Denote by ϕ_μ its characteristic function. Then $\phi_\mu \in L_1(\mathbb{R}^k)$ if and only if μ admits a continuous and bounded density f (on $\mathbb{R}^k$) given by*

$$f(x) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(ix \cdot \xi)\phi_\mu(-\xi)d\xi. \tag{4.2}$$

Proof. Assume that X has a density given by f_X. We, now, show that f given by Equation (4.2) coincide with f_X. The idea is to use Fubini theorem to exchange the order of integration of y and ξ but the lack of integrability prevents us to use it directly. For that purpose, we introduce a quantity on which it is possible to use Fubini's theorem and then see that it approximates the case of interest. Let

$$I_\varepsilon(x) = \frac{1}{(2\pi)^k} \iint_{\mathbb{R}^k \times \mathbb{R}^k} \exp(i(x-y)\cdot\xi)\exp\Big(-\varepsilon^2\frac{\xi^2}{2}\Big)d\mu(y)d\xi.$$

By integrating in y (implicitly using Fubini theorem) we get that

$$I_\varepsilon(x) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(ix\cdot\xi)\exp\Big(-\varepsilon^2\frac{\xi^2}{2}\Big)\phi_\mu(-\xi)d\xi$$

and then taking the limit for $\varepsilon \to 0$ and using dominated convergence theorem, we get

$$\lim_{\varepsilon\to 0} I_\varepsilon(x) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(ix\cdot\xi)\phi_\mu(-\xi)d\xi = f(x).$$

On the other side, by integrating first on the variable ξ, we get

$$\begin{aligned} I_\varepsilon(x) &= \frac{1}{(2\pi\varepsilon)^k} \int_{\mathbb{R}^k} \left(\int_{\mathbb{R}^k} \exp\Big(i\frac{1}{\varepsilon}(x-y)\cdot\varepsilon\xi\Big)\exp\Big(-\varepsilon^2\frac{\xi^2}{2}\Big)\varepsilon^k d\xi\right)d\mu(y) \\ &= \frac{1}{(\sqrt{2\pi}\varepsilon)^k} \int_{\mathbb{R}^k} \exp\Big(-\frac{\|x-y\|^2}{2\varepsilon}\Big)f_X(y)dy \end{aligned}$$

The quantity converges (in $L_1(\mathbb{R}^k)$) to $f_X(x)$ since the function ρ_ε defined by

$$\rho_\varepsilon(z) = \frac{1}{\varepsilon^k}\rho(z) \quad \text{where} \quad \rho(z) = \frac{1}{(2\pi)^{k/2}}\exp(-\frac{z^2}{2})$$

is a regularizing function (see Proposition 26). By unicity of the limit, $f_X = f$. The fact that X has a density implies $\phi_X \in L_1(\mathbb{R}^k)$ is obtained by considering $|\phi_X(\xi)| = \sqrt{\phi_X(\xi)\phi_{-X}(\xi)}$ and

$$\int_{-A}^{A} |\phi_X(\xi)| \leq \sqrt{\iint 2\frac{\sin(A(x-y))}{x-y} f_X(x) f_X(y) dx dy d\xi}$$

which is trivially upper bounded. For the other sense, the existence is the consequence of Equation (4.2) which gives the continuity of f_X by use of dominated convergence theorem. □

4.2.3 The characteristic function characterizes the law

The characterization of the law of a random variable is given by Theorem 3.

Proposition 9. *Let X and Y be two random vectors such that $\phi_X = \phi_Y$. Then, the distribution of X and the distribution of Y are equal.*

Proof. Let $Z \sim \mathcal{N}_k(0,1)$ be a gaussian random vector independent from X and Y. Let $\sigma > 0$ and the two random vectors $X_\sigma = X + \sigma Z$ and $Y_\sigma = Y + \sigma Z$ so that $\phi_{X_\sigma} = \phi_{Y_\sigma}$ (use Proposition 6). By Proposition 8, X_σ and Y_σ have continuous and bounded density. Now, using Theorem 3 we have that $X_\sigma \sim Y_\sigma$. Letting $\sigma \to 0$, we see that $X \sim Y$ by unicity of the limit for the convergence in distribution. □

4.3 Levy's theorem

Theorem 4. *Let $(F_n)_n$ be a sequence of distribution functions on the space $\mathbb{R}^k$ and for any $n \in \mathbb{N}$ let ϕ_n be the characteristic function of F_n. Suppose that*

$$\phi(\theta) := \lim \phi_n(\theta) \text{ exists for all } \theta \in \mathbb{R}^k.$$

Then, the following are equivalent.

i) The sequence $(X_n)_n$ is tight.

ii) The function ϕ is a characteristic function.

iii) The function ϕ is continuous at any θ in $\mathbb{R}^k$.

iv) The function ϕ is continuous at 0.

In particular, when one of these conditions is verified, there exists a distribution function F (hence there exists a random variable $X \sim F$) such that $\phi = \phi_F$ and

$$F_n \xrightarrow{(d)} F \quad (\text{or equivalently } X_n \xrightarrow{(d)} X).$$

Proof. We have $ii) \implies iii)$ from Proposition 6 and $iii) \implies iv)$ is obvious.
$i) \implies ii)$ By Helly Lemma (in Lemma 3), one can extract a sub-sequence n_k such that $F_{n_k} \xrightarrow{(d)} F$, where F is a distribution function (by the tightness of the sequence). By Lemma 1, we have that $\phi_{n_k} \longrightarrow \phi_F$ (pointwise). Obviously, one has to be careful about using Lemma 1 for Lipschitz function of complex values but one can always decompose $e^{i\theta X} = \cos(\theta X) + i\sin(\theta X)$ which are two real valued bounded Lipschitz functions. By unicity of the limit, we have $\phi = \phi_F$ and then ϕ is a characteristic function.
Proof of the last sentence We just showed the existence of the distribution function F. Now assume that F_n do not converge weakly to F. Then, there exists a point of continuity x of F (the set of points of continuity is never empty since the points of discontinuity are at most countable) and $\eta > 0$ such that there exists a sub-sequence $(n_i)_i$ such that

$$|F_{n_i}(x) - F(x)| \geq \eta.$$

By another use of Helly's lemma, one can find a sub-sequence of $(n_i)_i$ denoted $(n_{i_j})_j$ such that $F_{n_{i_j}} \xrightarrow{(d)} \widetilde{F}$ where $\widetilde{F}$ is a distribution function (by the tightness of the original sequence). Hence, $\phi_{n_{i_j}} \to \phi_{\widetilde{F}} = \phi_F$. By the uniqueness of the characteristic function (by Proposition 9), we have $\widetilde{F} = F$ and then $F_{n_{i_j}}(x) \to F(x)$ which is absurd.
$iv) \implies i)$ We first show the result in dimension 1 (k=1). Let $\varepsilon > 0$. The quantity $\phi_n(\theta) + \phi_n(-\theta)$ is real and bounded (by 2). By continuity of ϕ in 0, we can find $\delta > 0$ such that $\forall |\theta| < \delta$, $|1 - \phi(\theta)| < \varepsilon/4$ and

$$0 < \delta^{-1} \int_0^\delta (2 - \phi(\theta) - \phi(-\theta)) d\theta \leq \frac{\varepsilon}{2}.$$

Then by the (DOM) theorem (Theorem 14), $\exists n_0$ such that $\forall n \geq n_0$,

$$\delta^{-1} \int_0^\delta (2 - \phi_n(\theta) - \phi_n(-\theta)) d\theta \leq \varepsilon.$$

Then, first using Fubini theorem,

$$\varepsilon \geq \delta^{-1} \mathbb{E}\left[\int_{-\delta}^{\delta} (1 - e^{i\theta X_n}) d\theta\right] = 2\mathbb{E}\left[1 - \frac{\sin(\delta X_n)}{\delta X_n}\right] \geq 2\mathbb{E}\left[\mathbb{1}_{|X_n| > 2\delta^{-1}} \left(1 - \frac{1}{|\delta X_n|}\right)\right]$$
$$\geq \mathbb{E}\left[\mathbb{1}_{|X_n| > 2\delta^{-1}}\right] = \mathbb{P}\left(|X_n| > 2\delta^{-1}\right).$$

Since, the choice of δ is not depending on n, we have shown that the sequence $(X_n)_{n \geq n_0}$ is tight. But one can trivially add any finite sequence of random variables to a tight sequence and the resulting sequence keeps being tight.
For the general case, one has to replace the real valued quantity $\phi_n(\theta) + \phi_n(-\theta)$ by a new one. For $k = 2$, $f(\theta_1, \theta_2) = \phi_n(\theta_1, \theta_2) + \phi_n(\theta_1, -\theta_2) = \mathbb{E}\left[e^{i\theta_1 X_{n,1}} 2\cos(\theta_2 X_{n,2})\right]$. One has to define the real valued $g(\theta_1, \theta_2) = f(\theta_1, \theta_2) + f(-\theta_1, \theta_2)$ to replace the previous quantity. The arguments remain the same and are easily generalizable to any dimension. □

A obvious use of the previous theorem allows us to derive a usefull corollary.

Corollary 1 (Cramer-Wold device). *Let $(X_n)_n$ be a sequence of random variables in $\mathbb{R}^k$. Then*

$$X_n \xrightarrow{(d)} X \Leftrightarrow \forall t \in \mathbb{R}^k,\ t^T X_n \xrightarrow{(d)} t^T X$$

Proof. Exercice [ref section exercices] □

Example 2. *Let Z be a random vector of law $\mathcal{N}_k(\mu, \Sigma)$, [DEFINE THE DISTRIBUTION] then*

$$\phi_Z(\theta) = e^{i\theta^T \mu - \frac{1}{2}\theta^T \Sigma \theta}.$$

To see this, one can use the Cramer-Wold device and compute the characteristic function of $t^T Z$ for any $t \in \mathbb{R}^k$. The random variable $t^T Z$ is normal by definition and $\mathbb{E}\left[t^T Z\right] = t^T \mu$,

$$Var\left(t^T Z\right) = \mathbb{E}\left[(t^T Z - t^T \mu)^2\right] = \mathbb{E}\left[(t^T Z - t^T \mu)(t^T Z - t^T \mu)^T\right] = t^T \mathbb{E}\left[(Z - \mu)(Z - \mu)^T\right] t = t^T \Sigma t$$

Now using the result of Example 1, we have

$$\phi_Z(\theta) = \phi_{\theta^T \mu, \theta^T \Sigma \theta}(1) = \exp\left(i(\theta^T \mu) \times 1 - \frac{\theta^T \Sigma \theta \times 1^2}{2}\right)$$

4.4 Law of Large Numbers and Central Limit Theorem

4.4.1 The Central Limit Theorem

We use Theorem 4 to prove the classical weak version of the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT).

Theorem 5 (CLT). *Let $X_1, \ldots, X_n$ be i.i.d random variables (en $\mathbb{R}$) with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}\left[X_1^2\right] = \sigma^2$. Let $\overline{X}_n = n^{-1} \sum X_i$. Then, the sequence $\sqrt{n}\,\overline{X}_n$ converges in distribution towards $\mathcal{N}(0, \sigma^2)$.*

Proof. We use Levy's theorem. Let $\phi = \phi_{X_1}$. The existence of the two first derivative are given by Proposition 6 and $\phi'(0) = i\mathbb{E}[X_1] = 0$ as well as $\phi''(0) = i^2 \mathbb{E}\left[X_1^2\right] = -\sigma^2$. By independence, we see that

$$\mathbb{E}\left[e^{it\sqrt{n}\,\overline{X}_n}\right] = \phi^n\left(\frac{t}{\sqrt{n}}\right) = \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \underset{n \to +\infty}{\longrightarrow} e^{-\frac{t^2\sigma^2}{2}}.$$

Since the function $t \mapsto e^{-t^2\sigma^2/2}$ is continuous in 0 and is the characteristic function of $\mathcal{N}(0, \sigma^2)$, we have the conclusion. □

One can directly use the Cramer-Wold device to get the mutlidimensional version of the (CLT).

Theorem 6. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^k$, with $\mu = \mathbb{E}[X_1]$ and $\Sigma = \mathbb{E}\left[(X_1 - \mu)(X_1 - \mu)^T\right]$, we get that*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \xrightarrow{(d)} \mathcal{N}_k(0, \Sigma).$$

Proof. Use Cramer-Wold device with the fact that $\forall t \in \mathbb{R}^k$, the familly of $Y_i = (t^T X_i - t^T \mu)_i$ satisfies Theorem 5. □

4.4.2 The Law of Large Numbers

We show the weak version of law of large numbers. The naming *weak* comes from the fact that the convergence occurs in probability eventhough it is known to be true in the a.s. convergence under the same set of hypothesis. Nevertheless, a few more tools are needed for that purpose.

Proposition 10 (LLN). *Let $X_1, \dots, X_n$ be i.i.d random variables of characteristic function ϕ. Assume that $\phi'(0) = i\mu$ for a $\mu \in \mathbb{R}$, then $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$.*

Proof. Expanding ϕ, we get $\phi(t) = 1 + t\phi'(0) + o(t)$ when $t \to 0$. Then

$$\mathbb{E}\left[e^{it\overline{X}_n}\right] = \phi^n\left(\frac{t}{n}\right) = \left(1 + \frac{it\mu}{n} + o\left(\frac{1}{n}\right)\right)^n \underset{n\to+\infty}{\longrightarrow} e^{it\mu}$$

which is the characteristic function of a constant (equal to μ) random variable. Since the limit is constant, the convergence in distribution transfers to a convergence in probability (by Theorem 2). □

Exercice 14. *Show the mutlidimensional version of Proposition 10.*

4.5 Rare events theorem

Theorem 7 (Rare events)**.** *Let $(X_{n,j})_{1\le j\le M_n}$ be a family of independent Bernoulli random variables of parameter $p_{n,j}$. Assume that*

(i) M_n *is increasing and tends towards* $+\infty$.

(ii) $\sum_{j=1}^{M_n} p_{n,j} \underset{n\to+\infty}{\longrightarrow} \lambda > 0$.

(iii) $\max_{1\le j\le M_n} p_{n,j} \underset{n\to+\infty}{\longrightarrow} 0$.

Then, if $S_n = X_{n,1} + \cdots + X_{n,M_n}$, we have $S_n \xrightarrow{(d)} \mathcal{P}(\lambda)$ (the Poisson distribution of parameter λ).

Proof. By independence of the random variables $X_{n,j}$, we have that

$$\phi_{S_n}(t) = \prod_{j=1}^{M_n} \phi_{X_{n,j}}(t) = \prod_{j=1}^{M_n}(p_{n,j}e^{it} + 1 - p_{n,j}) = \prod_{j=1}^{M_n}(1 + p_{n,j}(e^{it} - 1)).$$

Let log be the principal determination of the complex logarithm (on $\mathbb{C}\backslash(-\infty, 0]$). Then, using Taylor's formula for the function $t \mapsto \log(1 + tz)$, we have that for any z such that $|z| < 1$,

$$\log(1 + z) = z - z^2 \int_0^1 (1 - u)\frac{1}{(1 + uz)^2}du.$$

Now take $z = e^{it} - 1$. By (iii), for n large enough, one has that $\max_{1\le j\le M_n} p_{n,j} \le 1/2$. So

$$\left|\sum_{j=1}^{M_n} p_{n,j}^2 z^2 \int_0^1 (1 - u)\frac{1}{(1 + up_{n,j}z)^2}du\right| \le \left(\max_{1\le j\le M_n} p_{n,j}\right)\sum_{j=1}^{M_n} p_{n,j}\int_0^1 (1 - u)\frac{1}{(1/2)^2}du \underset{n\to+\infty}{\longrightarrow} 0,$$

then $\log \phi_{S_n}(t)$ is well defined and

$$\sum_{j=1}^{M_n} \log(1 + p_{n,j}(e^{it} - 1)) \underset{n\to+\infty}{\longrightarrow} \lambda(e^{it} - 1).$$

This implies that $\phi_{S_n}(t) \to e^{\lambda(e^{it}-1)}$ which is the characteristic function of $\mathcal{P}(\lambda)$ and we conclude by using Levy's theorem. □

Chapter 5

Lindeberg-Feller theorem

The theorem of Lindeberg and Feller deals with the non-i.i.d. case in the Central Limit Theorem. It can also be used when the distribution of each variable depends on n, the number of observations.

Theorem 8 (Lindeberg-Feller)**.** *Let $(k_n)_n$ be a sequence of integers. For every $n \in \mathbb{N}$, we assume to have access to $(X_{n,1}, \dots, X_{n,k_n})$ a collection of independent random vectors (i.e. $\forall i, X_{n,i} \in \mathbb{R}^d$). Assume that*

1. $R_n := \sum_{i=1}^{k_n} \mathbb{E}\left[\|X_{n,i}\|^2 \mathbb{1}_{\|X_{n,i}\|>\varepsilon}\right] \underset{n\to+\infty}{\longrightarrow} 0, \quad \forall \varepsilon > 0.$

2. $\sum_{i=1}^{k_n} \mathrm{Cov}(X_{n,i}) \underset{n\to+\infty}{\longrightarrow} \Sigma$

Then $\sum_{i=1}^{k_n} X_{n,i} - \mathbb{E}[X_{n,i}] \underset{n\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0, \Sigma)$.

Proof. We divide the proof in **four** steps.

Step 1: Reduction to the real case Without any restriction of generality, we can assume (by a centering) $\mathbb{E}[X_{n,i}] = 0$. By the result of Cramer-Wold 1, it is sufficient to show that for all $t \in \mathbb{R}^d$,

$$t^T \sum_{i=1}^{k_n} X_{n,i} \underset{n\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}\left(0, t^T \Sigma t\right).$$

Let fix $t \in \mathbb{R}^d$. It is easy to see that the hypothesis of the theorem imply the same hypothesis for the random variables $t^T X_{n,i}$. Indeed,

$$\begin{aligned}\mathbb{E}\left[(t^T X_{n,i})^2 \mathbb{1}_{|t^T X_{n,i}|>\varepsilon}\right] &\le \mathbb{E}\left[\|t\|^2 \|X_{n,i}\|^2 \mathbb{1}_{\|t^T\|\|X_{n,i}\|>\varepsilon}\right] \\ &= \|t\|^2 \mathbb{E}\left[\|X_{n,i}\|^2 \mathbb{1}_{\|X_{n,i}\|>\frac{\varepsilon}{\|t\|}}\right] \longrightarrow 0\end{aligned}$$

and

$$\sum_{i=1}^{k_n} \mathbb{E}\left[(t^T X_{n,i})^2\right] = \sum_{i=1}^{k_n} \mathbb{E}\left[t^T X_{n,i} X_{n,i}^T t\right] = t^T \Big(\sum_{i=1}^{k_n} \mathrm{Cov}(X_{n,i})\Big) t \longrightarrow t^T \Sigma t.$$

Then, it is enough to show the theorem for real valued random variables only. For the rest of the proof, we assume that $\forall i,\ X_{n,i} \in \mathbb{R}$.

Step 2: Variance control We denote by $\sigma_{n,i}^2 = \mathbb{E}\left[X_{n,i}^2\right]$ and $\sigma_n^2 = \sum_{i=1}^{k_n} \sigma_{n,i}^2$, then, by assumption, σ_n^2 converges to a finite quantity that we denote σ^2. Furthermore,

$$\begin{aligned}\sup_{i=1,\dots,k_n} \sigma_{n,i}^2 &= \sup_{i=1,\dots,k_n} \left(\mathbb{E}\left[X_{n,i}^2 \mathbb{1}_{|X_{n,i}|\le\varepsilon}\right] + \mathbb{E}\left[X_{n,i}^2 \mathbb{1}_{|X_{n,i}|>\varepsilon}\right]\right) \\ &\le \varepsilon^2 + \sum_{i=1}^{k_n} \mathbb{E}\left[X_{n,i}^2 \mathbb{1}_{|X_{n,i}|>\varepsilon}\right] = \varepsilon^2 + R_n.\end{aligned}$$

Fix $\underline{\varepsilon_0 > 0}$ and $\varepsilon = \sqrt{\varepsilon_0/2}$. There exists $\underline{N_0}$ such that $\forall n \ge N_0,\ R_n \le \varepsilon_0/2$. Hence, $\sup_{i=1,\dots,k_n} \sigma_{n,i}^2$ tends to 0. By assumption, σ_n has a non-zero limit which implies that, $\forall \delta > 0, \exists n_0, \forall n \ge n_0, \forall i \in \{1, \dots, k_n\}$

$$|\sigma_{n,i}^2| \le \delta \sigma_n^2 \tag{5.1}$$

Step 3: An equivalence Let $S_n = \sum_{i=1}^{k_n} X_{n,i}$. We have to show that

$$\phi_{S_n}(t) \underset{n\to+\infty}{\longrightarrow} e^{-\frac{1}{2}t^2\sigma^2}. \tag{5.2}$$

We begin with showing that (5.2) is equivalent to

$$\sum_{i=1}^{k_n} \phi_{X_{n,i}}(t) - 1 \underset{n\to+\infty}{\longrightarrow} -\frac{1}{2}t^2\sigma^2. \tag{5.3}$$

For that purpose, we use the following lemma which is proved in Section 13.1.

Lemma 5. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be complex numbers such that $\forall i$, $|a_i| \le 1$ and $|b_i| \le 1$. Then*

$$|a_1 a_2 \dots a_n - b_1 b_2 \dots b_n| \le \sum_{i=1}^{n} |a_i - b_i|.$$

Using the previous lemma with the complex numbers $a_i = e^{\phi_{X_{n,i}}(t)-1}$ and $b_i = \phi_{X_{n,i}}(t)$, of modulus bounded by 1, we have

$$\begin{aligned} \left| e^{\sum_{i=1}^{k_n}(\phi_{X_{n,i}}(t)-1)} - \phi_{S_n}(t) \right| &\le \sum_{i=1}^{k_n} |e^{\phi_{X_{n,i}}(t)-1} - \phi_{X_{n,i}}(t)| \\ &\le \sum_{i=1}^{k_n} \frac{|\phi_{X_{n,i}}(t) - 1|^2}{2} \end{aligned} \tag{5.4}$$

where we used that for any $z \in \mathbb{C}$ such that $\Re(z) \le 0$, it holds that $|e^z - 1 - z| \le |z|^2/2$. See Lemma 21 for a proof of this fact. Using the Taylor-Young formula,

$$|\phi_{X_{n,i}}(t) - 1| = |\phi_{X_{n,i}}(t) - 1 - t\phi'_{X_{n,i}}(0)| = \left| \int_0^t (x-t)\phi''_{X_{n,i}}(x)dx \right| \le \frac{t^2}{2}\sigma_{n,i}^2.$$

where we used $|\phi''_{X_{n,i}}(x)| \le \mathbb{E}\left[X_{n,i}^2\right] = \sigma_{n,i}^2$. Then, plugging it in (5.4) and using (5.1), we finally show

$$\left| e^{\sum_{i=1}^{k_n}(\phi_{X_{n,i}}(t)-1)} - \phi_{S_n}(t) \right| \le \frac{t^4}{8} \sum_{i=1}^{k_n} \sigma_{n,i}^4 \le \frac{t^4}{8}\sigma_n^2 \delta$$

This shows that the left hand side quantity tends to 0 when n goes to infinity. Finally, by triangular inequality, we have showed (5.2) $\Leftrightarrow$ (5.3).

Finish It remains to show (5.3). By the mean value theorem, there exists $c_t \in [0,t]$ such that

$$\begin{aligned} \sum_{i=1}^{k_n} \phi_{X_{n,i}}(t) - 1 + \frac{t^2}{2}\sigma_n^2 &= \sum_{i=1}^{k_n} \phi_{X_{n,i}}(t) - \left(\phi_{X_{n,i}}(0) + t\phi'_{X_{n,i}}(0) + \frac{t^2}{2}\phi''_{X_{n,i}}(0)\right) \\ &= \sum_{i=1}^{k_n} \frac{t^2}{2}(\phi''_{X_{n,i}}(c_t) - \phi''_{X_{n,i}}(0)) \\ &= \sum_{i=1}^{k_n} \frac{t^2}{2}\mathbb{E}\left[-X_{n,i}^2(e^{ic_t X_{n,i}} - 1)\right] \\ &\le \frac{t^2}{2} \sum_{i=1}^{k_n} \mathbb{E}\left[X_{n,i}^2 |e^{ic_t X_{n,i}} - 1| \mathbb{1}_{|X_{n,i}| \le \varepsilon}\right] + t^2 \sum_{i=1}^{k_n} \mathbb{E}\left[X_{n,i}^2 \mathbb{1}_{|X_{n,i}| > \varepsilon}\right] \\ &\le \frac{t^2}{2} \sum_{i=1}^{k_n} c_t \varepsilon \sigma_{n,i}^2 + t^2 R_n \le \frac{t^3}{2}\sigma_n^2 \varepsilon + t^2 R_n \end{aligned}$$

Since this is true for every $\varepsilon > 0$ and that $\sigma_n^2 \longrightarrow \sigma^2$ and $R_n \longrightarrow 0$, we showed (5.3). By the use of Levy's theorem 4, on limiting characteristic function $t \mapsto e^{-t^2/2\sigma^2}$ of a centered normal with variance σ^2 (continuous at 0), we have finished the proof. □

5.0.1 Application to regression problems

Chapter 6

Dependent limit theorems

In this chapter we deal with the case of random variables that may be possibly weakly dependent. We assume that the random variables $(X_i)_i$ are centered (i.e. $\mathbb{E}[X_i] = 0$). If one wants to avoid assuming that condition, it will ba at the cost of assuming that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \underset{n\to+\infty}{\longrightarrow} \ell$$

for a $\ell \in \mathbb{R}$.

6.1 Weakly dependent laws of large numbers

6.1.1 Weak law of large numbers under dependence

Proposition 11. *Let $X_1, \ldots, X_n$ be real random variables such that $\forall i$, $\mathbb{E}[X_i] = 0$. Assume that*

- $\sum_i \mathit{Var}(X_i) = o(n^2)$
- *There exists $\phi : \mathbb{N} \to \mathbb{R}_+$ such that $\forall i, j$, $|\mathit{Cov}(X_i, X_j)| \leq \phi(|i-j|)$ and*

$$\frac{1}{n}\sum_{i=1}^{n}\phi(i) \underset{n\to+\infty}{\longrightarrow} 0$$

Then

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} 0.$$

Proof. By Chebyshev's inequality, it is sufficient to prove that $\mathrm{Var}(S_n) \to 0$.

$$\begin{aligned}
\mathrm{Var}(S_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{Cov}(X_i, X_j) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\phi(k) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\phi(k) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{2}{n}\sum_{k=1}^{n}\phi(k) = o(1)
\end{aligned}$$

□

Of course, one could replace the second condition of Proposition 11 by the stronger $\mathrm{Cov}(X_i, X_j) \to 0$ when $|i-j| \to \infty$. The first condition is trivially satisfied when the X_i's are identically distributed or when one can find $c > 0$ such that $\forall i$, $\mathrm{Var}(X_i) \leq c$.

6.1.2 Strong law of large numbers under dependence

One option to prove strong dependent law of large numbers is Lemma 38, at the cost of assuming a uniform bound by at integrable variable X.

Corollary 2. *Assume the hypothesis of Proposition 11 and* $\forall i$, $|X_i| \leq X$ *such that* $\mathbb{E}[X] < \infty$, *we obtain that*

$$S_n \xrightarrow{a.s.} 0.$$

It is also possible to prove a version of it using martingales techniques. The centering has to be handled carefully since, in general, the sum of dependent random variables does not satisfy the martingale axioms. We use the notation $\mathcal{F}_i$ for the filtration corresponding to $\sigma(X_1, \ldots, X_i)$.

Proposition 12. *Let* $X_1, \ldots, X_n$ *be real random variables such that* $\forall i$, $\mathbb{E}[X_i] = 0$. *Assume that there exists* $r \geq 0$ *such that for all* i, j *such that* $|i-j| \geq r$, X_i *and* X_j *are independent. Assume that*

- *For all* $i = 1, \ldots, n$ *and* $j = 1, \ldots, r$, $\mathbb{E}[Var(X_i|\mathcal{F}_{i-j})] \leq \sigma_i^2$.
- $\sum_i \sigma_i^2 < \infty$.

Then $\sum_{i=1}^n X_i$ *converges almost surely towards 0.*

Note that the assumptions of Proposition 12 include the assumptions of Proposition 11 when one apply it for the random variables $n^{-1}X_i$.

Proof. The proof uses the fact that a martingale bounded in L_2 is almost surely convergent. Let $Y_i = X_i - \mathbb{E}[X_i|\mathcal{F}_{i-1}]$ so that $M_t = \sum_{i=1}^t Y_i$ is a martingale.

$$\mathbb{E}\left[M_n^2\right] = \sum_{i=1}^n \mathbb{E}\left[(M_i - M_{i-1})^2\right] = \sum_{i=1}^n \mathbb{E}\left[Y_i^2\right] = \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}\left(X_i|\mathcal{F}_{i-1}\right)\right] \leq \sum_{i=1}^n \sigma_i^2$$

Then the martingale $(M_n)_n$ is bounded in L_2 and its limit exists and the convergence is almost sure. Now define $Z_i = \mathbb{E}[X_i|\mathcal{F}_{i-1}] - \mathbb{E}[X_i|\mathcal{F}_{i-2}]$. The sum $(N_n)_n$ of the random variables Z_i is again a martingale bounded in L_2 for the same kind of calculations. Then, identically, N_n converges almost surely. Following this scheme, we can write $\sum_i X_i$ as a sum of r martingales of the form

$$\sum_{i=1}^t \mathbb{E}[X_i|\mathcal{F}_{i-j}] - \mathbb{E}[X_i|\mathcal{F}_{i-j-1}]$$

that all converge almost surely. Then, $\sum_i X_i$ converges almost surely to a random variable X. Since the assumptions of Proposition 11 are fulfilled, the only possible limit is 0. □

Of course, one can imagine generalizations of the previous result when the resulting convergence for the martingales are of type 'bounded in L_1' only using first moments conditions. It is also possible to generalize Kolmogorov three series theorem in the case of weak dependence. Finally, the weak dependence condition of Proposition 12 does not have to be of uniform flavor and a bound depending on j is possible as long has one ask for the convergence of the series of variances.

6.2 Central Limit Theorems under dependence

In this section, we expose weak dependence central limit theorems using the ideas of Lindeberg-Feller theorem. This section follows the work of [5].

6.2.1 Bernstein blocks

Assume given a sequence of random variables $X_1, \ldots, X_n$, we decompose its sum into blocks of two different size. This is the so-called Berstein block technique. Let $(p_n)_n$ and $(q_n)_n$ be two sequences such that

$$p_n \underset{n\to+\infty}{\longrightarrow} +\infty, \quad q_n \underset{n\to+\infty}{\longrightarrow} +\infty, \quad q = o(p), \quad p = o(n).$$

We split $S_n = \sum_{i=1}^n X_i$ into blocks of different size. The benefit from this technique is to be able to make use of gaps (of size q_n) between blocks as well as the fact that the blocks of size q_n are too small to count in the final convergence.

$$S_n = \sum_{i=1}^k \varepsilon_i + \sum_{i=1}^{k+1} \nu_i = Z_k + Z'_{k+1},$$

where for $1 \leq i \leq k$,

$$\varepsilon_i = \sum_{(i-1)p+(i-1)q+1}^{ip+(i-1)q} X_j, \qquad \nu_i = \sum_{ip+(i-1)q+1}^{ip+iq} X_j \tag{6.1}$$

and $\nu_{k+1} = \sum_{k(p+q)+1}^{n} X_j$ where $p_n = p$, $q_n = q$ and $k = \lfloor n/(p+q) \rfloor$. In the following result, we encode the good assumptions to obtain that the part Z'_{k+1} does not influence the convergence.

Lemma 6. *Let $X_1, \dots, X_n$ be real random variables. Let $S_n = \sum_{i=1}^n X_i$ and $\sigma_n^2 = Var(S_n)$. Assume that for two sequences verifying (6.1), we have that*

1. $\frac{1}{\sigma_n^2} \mathbb{E}\left[Z'^2_{k+1}\right] \underset{n\to\infty}{\longrightarrow} 0,$
2. $C_{k,g,h}(t) := \sum_{j=2}^k \left| Cov\left(g\left(\frac{t}{\sigma_n}\sum_{i=1}^{j-1}\varepsilon_i\right), h\left(\frac{t}{\sigma_n}\varepsilon_j\right)\right)\right| \underset{n\to\infty}{\longrightarrow} 0,$ *for all $t \in \mathbb{R}$ and $g, h \in \{\cos, \sin\}$,*
3. $\frac{1}{\sigma_n^2}\sum_{i=1}^k \mathbb{E}\left[\varepsilon_i^2 \mathbb{1}_{|\varepsilon_i| \geq \varepsilon\sigma_n}\right] \underset{n\to\infty}{\longrightarrow} 0,$ *for all $\varepsilon > 0$,*
4. $\frac{1}{\sigma_n^2}\sum_{i=1}^k \mathbb{E}\left[\varepsilon_i^2\right] \underset{n\to\infty}{\longrightarrow} 1.$

Then, S_n/σ_n converges in distribution towards $\mathcal{N}(0,1)$.

Proof. Since $S_n/\sigma_n = Z_k/\sigma_n + Z'_{k+1}/\sigma_n$, assumption *1.* and Slutsky's lemma show that the limit in distribution of S_n/σ_n is the same as the limit of Z_k/σ_n. We follow the proof of Theorem 8 on the random variables ε_i. Assumptions *3.* and *4.* give an equivalent of (5.1) for the sequence $(\varepsilon_i)_i$ which is

$$\sup_i \sigma_{n,i}^2 \leq \delta\sigma_n^2,$$

where $\sigma_{n,i}^2 = \mathrm{Var}(\varepsilon_i)$. The challenging part is the one corresponding to **Step 3** of Theorem 8 and more particularly the first line of (5.4).

$$\begin{aligned}\left|e^{\sum_{i=1}^k(\phi_{\varepsilon_i/\sigma_n}(t)-1)} - \phi_{Z_k/\sigma_n}(t)\right| &\leq \left|e^{\sum_{i=1}^k(\phi_{\varepsilon_i/\sigma_n}(t)-1)} - \prod_{i=1}^k \phi_{\varepsilon_i/\sigma_n}(t)\right| + \left|\prod_{i=1}^k \phi_{\varepsilon_i/\sigma_n}(t) - \phi_{Z_k/\sigma_n}(t)\right| \\ &\leq \sum_{i=1}^k |e^{\phi_{\varepsilon_i/\sigma_n}(t)-1} - \phi_{\varepsilon_i/\sigma_n}(t)| + 4\max_{g,h\in\{\cos,\sin\}} C_{k,g,h}(t)\end{aligned}$$

where we used the fact that $e^{itx} = \cos(tx) + i\sin(tx)$ and a telescopic sum. The first term can be handled in the same way as in Theorem 8 whereas the second term tends to 0 by asumption. Finally, the convergence of $\sum_{i=1}^k(\phi_{\varepsilon_i/\sigma_n}(t)-1)$ is completely similar and we get that $Z_n/\sigma_n \xrightarrow{(d)} \mathcal{N}(0,1)$. □

[WRITE def1 and the proof of Proposition 1 of Doukhan]

Chapter 7

Concentration inequalities

In this chapter we derive an important class of results called concentration inequalities. They are a tool to control the deviation of a function of a certain number of independent random variables around its expected value. A concentration inequality is a result of the form

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \geq t\right) \leq g(t)$$

where the function g is a function depending on the distribution of Z.

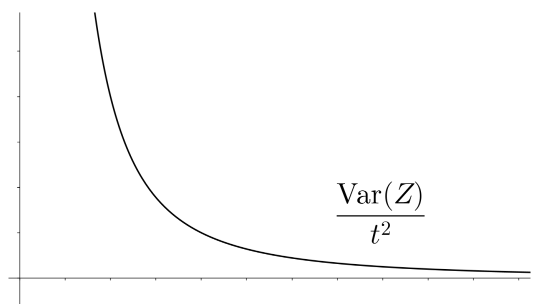


Figure 7.1: The concentration inequality of Bienaymé-Tchebychev

When $Z = Z_n := f_n(X_1, \dots, X_n)$ is function of independent random variables $X_1, \dots, X_n$, one includes the dependence in n in the deviation function so that

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \geq t\right) \leq g(n, t). \tag{7.1}$$

We expect to find a non-increasing function g with respect to its arguments n and t. The advantage of such results is that they permit to express statistical or probabilistic results valid for a fixed value of the number n of variables in the problem. It has to be expected that the concentration inequalities involve worse constants than in asymptotic theorems. Indeed, if we assume that Z_n converges to a limit variable Y, since the concentration inequalities (7.1) are valid for every n, and that the concentration of the asymptotic variable Y only verifies (7.1) in the limit sense, we logically get worse bounds. This chapter is highly inspired by the excellent [2].

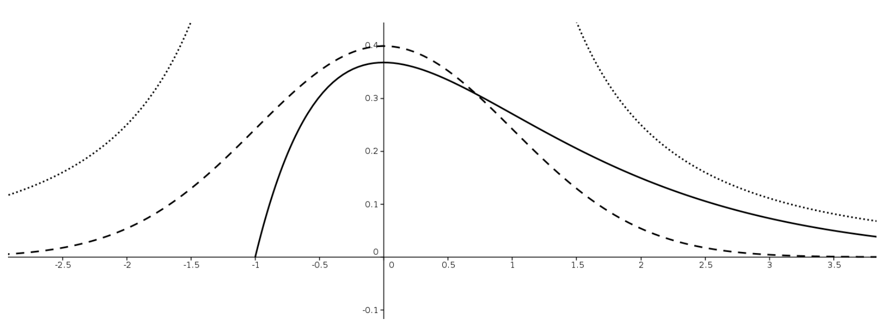


Figure 7.2: In solid line represents the distribution of a variable Z_n. The dotted line is a concentration inequality (here Bienaymé-Tchebychev). The dashed line represents the asymptotic distribution of the variable Z_n.

7.1 Chernoff Inequality

7.1.1 Basic principals

Here we show Markov's inequality and its direct consequences.

Proposition 13. *Let X be a real random variable. We assume that X is non-negative, then*

$$\forall t>0, \quad \mathbb{P}\left(X \geq t\right) \leq \frac{\mathbb{E}\left[X\right]}{t}.$$

Proof. We write $X = X\mathbb{1}_{X\geq t} + X\mathbb{1}_{X<t} \geq t\mathbb{1}_{X\geq t}$, hence taking the expectation we get the result. □

Exercice 15. *Show that a non-negative random variable X that can be written as $Yg(X)$ where g is a non-increasing function satisfy $\mathbb{P}\left(X \geq t\right) \leq \frac{\mathbb{E}[Y]g(t)}{t}$.*

Exercice 16. *Show that for any $p' > p \geq 1$, we have*

$$\mathbb{E}\left[|X|^p\mathbb{1}_{|X|\geq t}\right] \leq t^{1-\frac{p'}{p}}\mathbb{E}\left[|X|^{p'}\right]$$

A direct corollary of Markov inequality is the following so called Bienaymé-Tchebychev inequality.

Corollary 3. *For any real random variable X, we have that for any positive t, $\mathbb{P}\left(|X-\mathbb{E}\left[X\right]| \geq t\right) \leq \frac{Var(X)}{t^2}$.*

Proof. Apply the Markov's inequality for the non-negative random variable $(X-\mathbb{E}\left[X\right])^2$. □

The idea behind Bienaymé-Tchebychev inequality is somehow generic in the theory of concentration inequalities. The upcoming transformation of the random variable X replaces the transformation $X \to (X-\mathbb{E}\left[X\right])^2$ of the precedent proof the transform $x \to \exp(\lambda x)$ which depends on a parameter λ that is optimized at some step in the proof. The function

$$\lambda \geq 0 \mapsto \Psi_Z(\lambda) = \log \mathbb{E}\left[\exp(\lambda Z)\right]$$

is called the **Cramér-Chernoff** transform of Z. The dual function Ψ_Z^* is given by

$$\Psi_Z^*(t) = \sup_{\lambda\geq 0}(\lambda t - \Psi_Z(\lambda))$$

and is called **Fenchel-Legendre** transform. Following the path of the proof of Bienaymé-Tchebychev's inequality, we obtain (after optimization in λ) the following corollary.

Corollary 4. *For any real valued random variable Z, we have that*

$$\mathbb{P}\left(Z \geq t\right) \leq \exp\left(-\Psi_Z^*(t)\right)$$

for any $t > 0$.

Comments It is clear that $\Psi_Z(0) = 0$ which implies directly that $\Psi_Z^*(t) \geq 0$ as it is a suprema of a set containing 0. When $\mathbb{E}\left[Z\right]$ exists, Jensen's inequality implies that $\Psi_Z(t) \geq \lambda\mathbb{E}\left[Z\right]$. Hence, when $t \leq \mathbb{E}\left[Z\right]$, we have that $\lambda t - \Psi_Z(\lambda) \leq 0$ and $\Psi_Z^*(t) = 0$. This result is then empty when $t \leq \mathbb{E}\left[Z\right]$. For that specific reason, we will usually center the random variable Z (i.e. $\mathbb{E}\left[Z\right] = 0$ is assumed at the cost of changing Z into $Z - \mathbb{E}\left[Z\right]$). Furthermore, when $\mathbb{E}\left[Z\right] = 0$, $\lambda \leq 0$ and $t \geq 0$, another use of Jensen's inequality gives $\lambda t - \Psi_Z(\lambda) \leq 0$ and then

$$\Psi_Z^*(t) = \sup_{\lambda\in\mathbb{R}}(\lambda t - \Psi_Z(\lambda))$$

Proof. For any $\lambda \geq 0$, using Markov's inequality for the non-negative random variable $e^{\lambda Z}$ and by the monotonicity of the exponential,

$$\mathbb{P}\left(Z \geq t\right) \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda Z}\right] = e^{-(\lambda t - \Psi_Z(\lambda))}.$$

Now, using the fact that the probability on the left hand side is not depending on the parameter $\lambda \geq 0$, we finally have that

$$\mathbb{P}\left(Z \geq t\right) \leq \inf_{\lambda\geq 0} e^{-(\lambda t - \Psi_Z(\lambda))} = e^{-\Psi_Z^*(t)}.$$

□

7.1.2 Examples

Gaussian random variables Let Z be a gaussian $\mathcal{N}\left(0,\sigma^2\right)$ random variable. Since $\mathbb{E}\left[e^{\lambda Z}\right]=e^{\lambda^2\sigma^2/2}$, $\Psi_Z(\lambda)=\lambda^2\sigma^2/2$. Then

$$\Psi_Z^*(t)=\sup_{\lambda\in\mathbb{R}}(\lambda t-\frac{\lambda^2\sigma^2}{2})=\frac{t^2}{2\sigma^2}$$

and as expected, one has $\mathbb{P}\left(Z\geq t\right)\leq e^{-t^2/2\sigma^2}$.

Poisson random variables Let Y be a Poisson $\mathcal{P}(\nu)$ random variable and define $Z=Y-\nu$. The moment generating function is given by

$$\mathbb{E}\left[e^{\lambda Z}\right]=e^{-\lambda\nu}e^{-\nu}\sum_{k=0}^{\infty}\frac{(e^{\lambda}\nu)^k}{k!}=e^{-\lambda\nu-\nu}e^{\nu e^{\lambda}},$$

then $\Psi_Z(\lambda)=\nu(e^{\lambda}-\lambda-1)$. Let $f_t(\lambda)=\lambda t-\nu(e^{\lambda}-\lambda-1)$, then $f_t'(\lambda)=t-\nu(e^{\lambda}-1)$ and the maximum of f_t is attained at $\lambda=\log(1+t/\nu)$. This gives

$$\Psi_Z^*(t)=\nu h(t/\nu)\quad\text{where}\quad h(x)=(1+x)\log(1+x)-x.$$

Since $h(x)\underset{x\to+\infty}{\sim}x\log(x)$, $\Psi_Z^*(t)\underset{t\to+\infty}{\sim}t\log(t/\nu)\sim t\log(t)$ and then $\mathbb{P}\left(Y-\nu\geq t\right)=\underset{t\to+\infty}{O}\left(e^{-t\log(t)}\right)$. With extra calculation, one can easily prove that

$$h(x)\geq\frac{x^2}{2(1+\frac{u}{3})},$$

which actually shows that the Poisson random variables have a sub-Gamma tail in the sense of Definition 5 below.

Sub-Gamma random variables See Proposition 14.

7.1.3 Sub-Gaussian and sub-Gamma random variables

Definition 5. *We say that a random variable X is*

- *a* ***sub-Gaussian*** *random variable of constant $\nu>0$ if $\forall\lambda\in\mathbb{R}$, $\Psi_X(\lambda)\leq\lambda^2\nu/2$. We denote $X\in\mathcal{G}(\nu)$.*
- *a* ***sub-Gamma*** *random variable to the right, of constant $\nu>0$ and $c>0$ if*

$$\Psi_X(\lambda)\leq\frac{\lambda^2\nu}{2(1-2c\lambda)}\quad\text{for any } 0<\lambda<1/c.$$

We denote $X\in\Gamma_+(\nu,c)$. If $-X$ is sub-Gamma to the right, we say that $X\in\Gamma_-(\nu,c)$. We finally note $\Gamma(\nu,c)=\Gamma_+(\nu,c)\cap\Gamma_-(\nu,c)$.

For equivalent definitions/characterization of sub-Gaussian and sub-Gamma random variables, one can take a look at the first chapter of [2]. Of course, the vocabulary is relevant as seen in the following example.

Example 3. *A gaussian $\mathcal{N}\left(0,\sigma^2\right)$ random variable is sub-gaussian $\mathcal{G}(\sigma^2)$.*

Example 4. *A Gamma random variable Y of parameters (a,b) is sub-Gamma $\Gamma(ab^2,b)$. Indeed, $\mathbb{E}\left[Y\right]=ab$ and $Var(Y)=ab^2$. Let $X=Y-ab$ then for any $\lambda<1/b$,*

$$\mathbb{E}\left[e^{\lambda X}\right]=\int_0^{+\infty}e^{\lambda(y-ab)}\frac{y^{a-1}e^{-\frac{y}{b}}}{\Gamma(a)b^a}dy=e^{-\lambda ab}(1-\lambda b)^{-a}$$

and $\forall\lambda<1/b$, $\Psi_X(\lambda)=-\lambda ab-a\log(1-\lambda b)$. But since, $\log(1-u)-u\leq u^2/(2(1-u))$, we have

$$\forall\lambda\in(0,1/b),\quad\Psi_X(\lambda)\leq\frac{\lambda^2ab^2}{2(1-\lambda b)}.$$

For $\lambda<0$, which correspond to computing the Legendre transform for $-X$, using $-\log(1-u-u\leq u^2/2)$, we get

$$\Psi_X(\lambda)\leq\frac{a\lambda^2b^2}{2}$$

which gives that $X_-\in\mathcal{G}(ab^2)\subset\Gamma_+(ab^2,0)\subset\Gamma_+(ab^2,b)$. Then $X\in\Gamma(ab^2,b)$. It is interesting to see that the two tails of a Gamma random variable are unbalanced. The right part is sub-Gamma whereas the left tail is actually sub-Gaussian. In some cases, the behaviors of the tails on the left and on the right are different and one may study them separately.
The tail of the sub-Gamma concentration is slightly different from the concentration of sub-Gaussian random variables. The precise statement is as follows.

Proposition 14. *Let $X \in \Gamma(\nu, c)$ then for all $t > 0$,*

$$\mathbb{P}\left(X > \sqrt{2\nu t} + ct\right) \leq e^{-t} \qquad \mathbb{P}\left(-X > \sqrt{2\nu t} + ct\right) \leq e^{-t}$$

Proof. Since $\Psi_X(\lambda) \leq \frac{\nu\lambda^2}{2(1-c\lambda)}$,

$$\Psi_X(t) \geq \sup_{\lambda \in (0,1/c)} \left(t\lambda - \frac{\lambda^2 \nu}{2(1-c\lambda)}\right) = \frac{\nu}{c^2} g\left(\frac{ct}{\nu}\right)$$

where $g(u) = 1 + u - \sqrt{1+2u}$ for $u \geq 0$. Then

$$\mathbb{P}\left(X \geq t\right) \leq \exp\left(-\frac{\nu}{c^2} g\left(\frac{ct}{\nu}\right)\right)$$

but since $g^{-1}(u) = u + \sqrt{2u}$, one has directly $\mathbb{P}\left(X > \sqrt{2\nu t} + ct\right)$. The left tail is handled in the same way. □

The following result deals with the inverse of the Fenchel-Legendre transform.

Lemma 7. *Let ψ be a convex function such that $\psi(0) = \psi'(0) = 0$ that we assume differentiable on $[0, b)$ for $0 < b \leq +\infty$. For any $T \geq 0$, we define,*

$$\psi^*(t) = \sup_{\lambda \in [0,b)} (\lambda t - \psi(\lambda)).$$

Then ψ^* is positive, increasing and convex on $(0, +\infty)$, is such that $\psi^*(0) = 0$ and

$$\psi^{*-1}(y) = \inf_{\lambda \in (0,b)} \left[\frac{y + \psi(\lambda)}{\lambda}\right].$$

Proof. As a direct consequence of the assumptions, ψ is a non-decreasing function and then is non-negative on $[0, b)$. This triggers that $\psi^*(0)$ is a supremum of non-positive values where 0 is among them. This shows that $\psi^*(0) = 0$. As a supremum of convex and non-decreasing functions ψ^* is convex and non-decreasing. Hence ψ^* is non-negative. Now assume that there exists $t > 0$, such that $\psi^*(t) = 0$, then for any $\lambda \in (0, b)$, $\psi(\lambda) \geq \lambda t$. But, then $\psi'(0) \geq t > 0$ which is absurd. This shows directly that ψ^* is also increasing. Let

$$u = \inf_{\lambda \in (0,b)} \left[\frac{y + \psi(\lambda)}{\lambda}\right]$$

then for any $t \geq 0$,

$$u \geq t \quad \Leftrightarrow \quad \forall \lambda \in (0, b),\ \frac{y + \psi(\lambda)}{\lambda} \geq t \quad \Leftrightarrow \quad \forall \lambda \in (0, b),\ y \geq \lambda t - \psi(\lambda) \quad \Leftrightarrow \quad y \geq \psi^*(t).$$

This equivalence shows that $u = \psi^{*-1}(y)$ in the generalized inverse framework but since the actual inverse of ψ^* exists (since it is a continuous increasing function on $(0, b)$) it coincide with the regular notion of inverse. □

Proposition 15. *Let $Z_1, \ldots, Z_N$ be real valued random variables such that for all $\lambda \in (0, b)$,*

$$\forall i,\ \Psi_{Z_i}(\lambda) \leq \psi(\lambda)$$

where ψ is convex differentiable and such that $\psi(0) = \psi'(0) = 0$. Then

$$\mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right] \leq \psi^{*-1}(\log N).$$

Proof. For any $\lambda \in (0, b)$,

$$\exp\left(\lambda \mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right]\right) \leq \sum_{i=1}^{N} \mathbb{E}\left[\exp(\lambda Z_i)\right] \leq N \exp(\psi(\lambda)),$$

which is equivalent to writing that $\lambda \mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right] - \psi(\lambda) \leq \log N$ or by optimization in λ,

$$\mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right] \leq \inf_{\lambda \in (0,b)} \left[\frac{\log N + \psi(\lambda)}{\lambda}\right].$$

We end by using Lemma 7. □

Using Proposition 15, one can directly derive a bound for the expectation of the maximum of sub-Gamma random variables. If $Z_i \in \Gamma(\nu, c)$,

$$\mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right] \leq \sqrt{2\nu \log N} + c \log N.$$

7.1.4 Hoeffding inequality

We begin with the concentration of a single bounded random variable.

Lemma 8 (Hoeffding lemma). *Let Y be a random variable with $\mathbb{E}[Y] = 0$ and such that $Y \in [a, b]$ and let $\Psi_Y(\lambda) = \log \mathbb{E}[\exp(\lambda Y)]$. Then, $\Psi''(\lambda) \leq \frac{(b-a)^2}{4}$ and $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$.*

Proof. Since $|Y - (a+b)/2| \leq (b-a)/2$, $\mathrm{Var}(Y) = \mathrm{Var}(Y - (a+b)/2) \leq (b-a)^2/4$. For any $\lambda > 0$, we define a modification P_λ of the distribution P of Y by $dP_\lambda(x) = e^{-\Psi_Y(\lambda)} e^{\lambda x} dP(x)$. Then, since the support of P_λ is also $[a, b]$, we have that, for $Z_\lambda \sim P_\lambda$, $\mathrm{Var}(Z_\lambda) \leq (b - a^2/2)$. But, immediate computations give

$$\Psi''_Y(\lambda) = e^{-\Psi_Y(\lambda)} \mathbb{E}\left[Y^2 e^{\lambda Y}\right] - e^{-2\Psi_Y(\lambda)} (\mathbb{E}\left[Y e^{\lambda Y}\right])^2 = \mathbb{E}\left[Z_\lambda^2\right] - \mathbb{E}\left[Z_\lambda\right]^2 = \mathrm{Var}(Z_\lambda) \leq \frac{(b-a)^2}{4}.$$

By integration and the fact that $\Psi'_Y(\lambda) = \mathbb{E}[Y] = 0$ and also that $\Psi_Y(0) = 0$, we get

$$\Psi'_Y(\lambda) \leq \frac{\lambda(b-a)^2}{4} \qquad \text{and} \qquad \Psi_Y(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

This concludes the proof. □

This inequality applies directly in the context of sums of Rademacher variables. Indeed, if $S = \sum_{i=1}^n \varepsilon_i a_i$ then

$$\mathbb{E}\left[e^{\lambda S}\right] \leq e^{\frac{\lambda^2}{8} \sum_{i=1}^n a_i^2}. \tag{7.2}$$

A natural application of the precedent result is the so called Hoeffding inequality given in the following theorem.

Theorem 9 (Hoeffding inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that for any i, $X_i \in [a_i, b_i]$. Let $S = \sum_{i=1}^n X_i - \mathbb{E}[X_i]$. Then $S \in \mathcal{G}(\sum_{i=1}^n (b_i - a_i)^2/4)$ and*

$$\forall t \geq 0, \quad \mathbb{P}(S \geq t) \leq \exp\left(-\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof. Obviously,

$$\mathbb{E}\left[e^{\lambda S}\right] \leq \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq \prod_{i=1}^n e^{\frac{\lambda^2(b_i-a_i)^2}{8}} = e^{\frac{\lambda^2 \sum_{i=1}^n (b_i-a_i)^2}{8}}$$

□

We only derived the result as an upper bound on the tail probability on the right tail but the symmetrical nature of the hypothesis give the exact same bound for the left tail. There are two generalization of the previous result that benefit from a different manner to control the Laplace transform of the random variables. The independence is obviously crucial in the arguments of the proofs. The first result gives that the tail concentration is of the form of a Poisson tail whereas the second ensures that the tail is sub-Gamma.

Proposition 16 (Benett inequality). *Let $b > 0$ and $\phi(u) = e^u - u - 1$ for $u \in \mathbb{R}$. Let $X_1, \ldots, X_n$ be independent random variables such that $\forall i, X_i \leq b$. Define $v = \sum_{i=1}^n \mathbb{E}\left[X_i^2\right]$ and $S = \sum_{i=1}^n X_i - \mathbb{E}[X_i]$. Then*

$$\Psi_S(\lambda) \leq \frac{v}{b^2} \phi(b\lambda).$$

Consequently, $\forall t \geq 0$ we have $\mathbb{P}(S \geq t) \leq \exp(-(v/b^2) h(tb/v))$ where $h(x) = (1+x)\log(1+x) - x$ for any $x > 0$.

The attentive reader has noticed that the tail bound is exactly of the nature of a Poisson concentration as seen in Section 7.1.2.

Proof. We assume $b = 1$ at the price of changing the random variables X_i in X_i/b. Notice that the function $u \mapsto \phi(u)/u^2$ is decreasing on R then, using that $X_i \leq 1$,

$$e^{\lambda X_i} - X_i - 1 \leq X_i^2(e^\lambda - \lambda - 1),$$

which induces $\mathbb{E}\left[e^{\lambda X_i}\right] - \mathbb{E}\left[X_i\right] - 1 \leq \mathbb{E}\left[X_i^2\right]\phi(\lambda)$. Then,

$$\begin{aligned}\Psi_S(\lambda) &= \sum_{i=1}^n (\Psi_{X_i}(\lambda) - \lambda\mathbb{E}\left[X_i\right]) \\ &\leq \sum_{i=1}^n \log\left(1 + \lambda\mathbb{E}\left[X_i\right] + \mathbb{E}\left[X_i^2\right]\phi(\lambda)\right) - \lambda\mathbb{E}\left[X_i\right] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\phi(\lambda) = v\phi(\lambda).\end{aligned}$$

□

A even stronger result is the following.

Proposition 17 (Bernstein inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that there exist $c > 0$ and $v > 0$ such that $\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq v$ and*

$$\forall q \geq 3, \quad \sum_{i=1}^n \mathbb{E}\left[(X_i)_+^q\right] \leq \frac{q!}{2} v c^{q-2}$$

where $x_+ = \max(x, 0)$. Then, denoting $S = \sum_{i=1}^n X_i - \mathbb{E}\left[X_i\right]$, we have that for all $\lambda \in (0, 1/c)$ and $t > 0$,

$$\Psi_S(\lambda) \leq \frac{v\lambda^2}{2(1-c\lambda)} \qquad \text{and} \qquad \Psi^*(t) \geq \frac{v}{c^2} g\left(\frac{ct}{v}\right)$$

where $g(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$.

Then the concentration is sub-Gamma on the right. Obviously, one has to keep in mind that this result as the previous one is not symmetric and then only holds for the right tail of the distribution of the sum. If one wants to get symmetric concentration, the conditions have to be assumed on both sides of the distributions of the X_i.

Proof. We use another time the notation $\phi(u) = e^u - u - 1$. For $u \leq 0$, $\phi(u) \leq u^2/2$. Then, for $\lambda > 0$,

$$\phi(\lambda X_i) \leq \phi(\lambda(X_i)_-)\mathbb{1}_{X_i<0} + \phi(\lambda(X_i)_+)\mathbb{1}_{X_i\geq 0} \leq \frac{\lambda(X_i)_-^2}{2} + \sum_{q=2}^{+\infty}\frac{\lambda^q (X_i)_+^q}{q!} = \frac{\lambda(X_i)^2}{2} + \sum_{q=3}^{+\infty}\frac{\lambda^q (X_i)_+^q}{q!}$$

from which we deduce $\sum_{i=1}^n \mathbb{E}\left[\phi(\lambda X_i)\right] \leq \frac{v}{2}\sum_{q=2}^{\infty}\lambda^q c^{q-2}$. Finally,

$$\begin{aligned}\Psi_S(\lambda) &= \sum_{i=1}^n \log\mathbb{E}\left[e^{\lambda X_i}\right] - \lambda\mathbb{E}\left[X_i\right] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] - 1 - \lambda\mathbb{E}\left[X_i\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[\phi(\lambda X_i)\right] \\ &\leq \frac{v\lambda^2}{2}\sum_{q=0}^{\infty}(\lambda c)^q = \frac{v\lambda^2}{2(1-c\lambda)}.\end{aligned}$$

The rest of the proof is very identical to the proof of Proposition 14. □

Part II

Statistics

Chapter 8

Convergence of empirical processes

8.1 Introduction

The simple convergences given by LLN and CLT,

$$\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X] \quad \text{or} \quad \sqrt{n}(\overline{X}_n - \mathbb{E}[X]) \xrightarrow{(d)} \mathcal{N}\left(0, \sigma^2\right)$$

gives that for any **fixed** function f in a set of functions $\mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} \mathbb{E}[f(X)] \quad \text{and} \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(X)]) \xrightarrow{(d)} \mathcal{N}\left(0, \sigma_f^2\right).$$

Many statistical contexts need to deal with the case when the function f is actually random and possibly dependent on the values of the random variables $X_1, \dots, X_n$. This case comes naturally when one needs a control of the empirical quantity $\frac{1}{n}\sum_{i=1}^{n} \hat{f}(X_i)$ for an estimator $\hat{f}$ drawn on the sample.

Measurability of the sup In the sequel, we will be interested in the behavior (in a large sense) of processes of the form $\sup_{f\in\mathcal{F}} f(X)$. Assume that one wants to give a precise meaning to $\mathbb{E}\left[\sup_{f\in\mathcal{F}} f(X)\right]$. The attentive reader noticed that this may pose a measurability problem. Indeed, in general, it is not possible to prove that $\sup_{f\in\mathcal{F}} f(X)$ is measurable. Nevertheless, there are, at least, two strategies to overcome this issue. First, one can define

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} f(X)\right] = \sup\left\{\sup_{f\in\mathcal{G}} f(X) : \text{ with } \mathcal{G} \text{ finite}\right\}. \tag{8.1}$$

Hence, one can always assume the suprema taken over finite sets (which define measurable objects). Secondly, one can define an improper notion of espectation $\mathbb{E}^*$ using the outer measure/outer integrals concept. For more information on that subject, one can relate the following notion with (15.2). For a map $Z : \Omega \to \mathbb{R}$, we define

$$\mathbb{E}^*[Z] = \inf\{\mathbb{E}[U] : \text{ with } U \geq Z \text{ and } U : \Omega \to [-\infty, +\infty] \text{ is measurable}\}. \tag{8.2}$$

Then one can define $\mathbb{E}^*\left[\sup_{f\in\mathcal{F}} f(X)\right]$ in this manner. Of course, if the set $\mathcal{F}$ is finite the two definition coincide with the regular notion of expectation on $\sup_{f\in\mathcal{F}} f(X)$ that is now clearly measurable. We do not specify which of these notions are of interest for us since, the conditions that we assume in the forthcoming theorems impose continuity a.s. of the process and separability of $\mathcal{F}$. These two conditions are sufficient to have that the supremum is actually measurable. In particular, it reduces to the study of a continuous random process defined on a Polish space (even though the completeness is not present).

8.2 Examples

8.2.1 Education vs Employment

In our model a population of individuals $X_1 = (Y_1, Z_1), \dots, X_n = (Y_n, Z_n)$ is such that $Y_i \in \{0, 1\}$ represents the fact for individual i to be employed (value 1) and $Z_i \in \mathbb{R}$ represents the level of education. We are interested in understanding the relation of dependence between education and employment summarized in the following function,

$$F_0(z) = \mathbb{P}(Y = 1 | Z = z).$$

A natural hypothesis to impose on the function F_0 is to be non-decreasing in z as (normally) a higher level of education gives more access to employment. Let

$$\Lambda_1 = \{F : \mathbb{R} \to [0,1],\ F \text{ is non decreasing}\}$$

a set of functions that satisfy the same conditions than F_0. A natural estimator for F_0 is the maximum likelihood estimator defined as

$$\hat{F}_n = \text{argmax}_{f\in\Lambda_1}\left\{\sum_{i=1}^n \big(Y_i \log F(Z_i) + (1-Y_i)\log(1-F(Z_i))\big)\right\}$$

Denoting by Q the distribution of the random variable Z. A measure of the quality of this estimator can be given by

$$\|\hat{F}_n - F_0\|_Q = \left(\int (\hat{F}_n(z) - F_0(z))^2 dQ(z)\right)^{1/2}.$$

The tools developed later in this chapter can be applied to get $\|\hat{F}_n - F_0\|_Q = O_P(n^{-1/3})$.
One may choose to impose some extra assumptions on the objective function by defining

$$\Lambda_2 = \left\{F : \mathbb{R} \to [0,1],\ 0 \le \frac{dF}{dz}(z) \le M,\ F \text{ is concave.}\right\}$$

In this context, it will be possible to show later in this chapter that $\|\hat{F}_n - F_0\|_Q = O_P(n^{-2/5})$. Finally, if one is interested in a parametric case and defines

$$\Lambda_3 = \left\{F : \mathbb{R} \to [0,1],\ F(z) = F_0(\theta z),\ \theta \in \mathbb{R} \text{ and } F_0(x) = \frac{e^x}{1+e^x}\right\}.$$

In this case, $\|\hat{F}_n - F_0\|_Q \le C|\hat{\theta}_n - \theta_0| = O_P(n^{-1/2})$.

8.2.2 Theoretical convergence of maximum likelihood estimators for densities

Assume that we are provided with a set of densities (with respect to a given measure μ)

$$\{p_\theta :\ \theta \in \Theta\}$$

to which belongs a density p_{θ_0}. The statistician is provided with a sample $X_1, \dots, X_n$ of common distribution p_{θ_0}. A suitable notion of distance for this problem is the so-called **Hellinger** distance h given by

$$h(p,q) = \left(\frac{1}{2}\int (p^{1/2} - q^{1/2})^2 d\mu\right)^{1/2}.$$

This distance is controlled by the Kullback-Leibler divergence (which is not properly a distance) K that is defined by

$$K(p,q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x) d\mu(x).$$

Note that the integrand is continuous (and takes the value 0) on the frontier of the support of p, hence no problems of integration occur in this case. Obviously, the $K(p,q) = +\infty$ if q is not absolutely continuous with respect to p.

Proposition 18. *We have that $K(p,q) \ge 0$ and that $h^2(p,q) \le \frac{1}{2}K(p,q)$.*

Proof. At the cost of reducing the set of integration to the support of p, we can assume that $p(x) > 0$ and $q(x) > 0$. A simple function study shows that $\forall v > 0$, we have

$$\log(v) \le v - 1 \quad \text{and} \quad \frac{1}{2}\log(v) \le v^{1/2} - 1$$

Hence,

$$K(p,q) = \int \log\left(\frac{p}{q}\right) p d\mu \ge \int \left(\frac{q}{p} - 1\right) p d\mu = \int q d\mu - \int p d\mu = 1 - 1 = 0$$

$$\frac{1}{2}K(p,q) = \int \frac{1}{2}\log\left(\frac{p}{q}\right) p d\mu \ge \int \left(1 - \frac{q^{1/2}}{p^{1/2}}\right) p d\mu = 1 - \int p^{1/2}q^{1/2} d\mu = \frac{1}{2}\left(\int p d\mu + \int q d\mu - \int 2p^{1/2}q^{1/2} d\mu\right) = h^2(p,q)$$

□

The maximum likelihood estimator is given by

$$p_{\hat{\theta}_n} = \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right)$$

where the right hand side can be interpreted as the empirical version of the Kullback-Leibler divergence. By definition of the estimator, we have

$$0 \geq \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_{\hat{\theta}_n}(X_i)}\right) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_{\hat{\theta}_n}(X_i)}\right) - K(p_{\theta_0}, p_{\hat{\theta}_n}) + K(p_{\theta_0}, p_{\hat{\theta}_n}).$$

Then

$$K(p_{\theta_0}, p_{\hat{\theta}_n}) \leq \left|\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_{\hat{\theta}_n}(X_i)}\right) - K(p_{\theta_0}, p_{\hat{\theta}_n})\right| \leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right) - K(p_{\theta_0}, p_\theta)\right|.$$

But one already know that for any fixed $\theta \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right) - K(p_{\theta_0}, p_\theta) = O_P(n^{-1/2}).$$

Finally, one can see that if one is able to derive a uniform type of central limit theorem, one will be able to give the order of magnitude of the convergence of $K(p_{\theta_0}, p_{\hat{\theta}_n})$ towards 0.

8.3 Metric entropy, covering and ε-nets

8.3.1 Covering numbers

We begin with the definition of the metric entropy in a general pseudo-metric space $\mathbb{D}$. The space is endowed with a pseudo-distance d (i.e the only axiom of a distance that is not verified by d is $d(x, y) = 0 \implies x = y$). In the following, we denote by $B_d(x, \varepsilon)$ the open ball centered at x and of radius $\varepsilon > 0$.

Definition 6. *Let $(\mathbb{D}, d)$ be a pseudo metric space.*

- *A* ***covering of radius*** *ε of a set A in the metric space $\mathbb{D}$ is a set $\mathcal{C}$ defined as a finite union of balls of the form $B_d(x, \varepsilon)$ such that $\mathcal{C}$ contains A. The elements $x \in \mathbb{D}$ do not necessarily belong to A.*
- *The set of coverings of A is denoted* ***Cov***(A).
- *For a covering $\mathcal{C}$ of A, we denote by* ***Centers***$(\mathcal{C})$ *the set of the centers x of the balls used in the covering $\mathcal{C}$.*

We define the ***covering number*** *$\mathcal{N}(\varepsilon, A, d)$ as the minimal number of balls needed to cover A:*

$$\mathcal{N}(\varepsilon, A, d) = \min_{\mathcal{C}\in Cov(A)} |Centers(\mathcal{C})|.$$

Note that the min *is a priori an* inf *but the number of elements in $Centers(\mathcal{C})$ is an integer and since the infimum is taken over a subset of natural numbers, this is a minimum. The quantity $H(\varepsilon, A, d) = \log \mathcal{N}(\varepsilon, A, d)$ is the ε-**entropy** of the set A. Finally, we say that the set A is **totally bounded** is the ε-entropy $H(\varepsilon, A, d)$ is finite for every $\varepsilon > 0$.*

Since we are interested in sets that are totally bounded, it is not important to assume that the centers belong to A or not. Indeed, if a covering $\mathcal{C} = \cup_i B_d(x_i, \varepsilon)$ exists, it is always possible to find another covering $\cup_i B_d(x_i', 2\varepsilon)$ where $x_i' \in A$. In the literature, a covering such that the x_i belong to A is called an internal covering and is called an external covering in the opposite case.

Entropy of a set of functions When the metric space is $L_p(\mathbb{R})$, we denote by $H(\varepsilon, \mathcal{F}, Q) := H(\varepsilon, \mathcal{F}, \|\cdot\|_{p,Q})$ the entropy of the set $\mathcal{F}$ with respect to the metric

$$d(f, g) = \|f - g\|_{p,Q} = \left(\int_{\mathbb{R}} |f - g|^p dQ\right)^{1/p}.$$

Of course, as in Definition 6, the set $\mathcal{F}$ is included in the ambient metric space which is $L_p(Q)$ in this case. We denote by $H_\infty(\varepsilon, \mathcal{F})$ the ε-**entropy for the infinite norm** $\|\cdot\|_\infty$.

Definition 7. *We denote by $\mathcal{N}_{p,B}(\varepsilon, \mathcal{F}, Q)$ the minimal number N such that there exists couples $(f_i^L, f_i^R)_{i=1}^N$ of elements of $L_p(Q)$ such that*

- *For all i, $\|f_i^L - f_i^R\|_{p,Q} \leq \varepsilon$.*
- *For all $f \in \mathcal{F}$, there exists $i \in \{1, \dots, N\}$ such that $f_i^L \leq f \leq f_i^R$.*

*The value of $H_{p,B}(\varepsilon, \mathcal{F}, Q) = \log \mathcal{N}_{p,B}(\varepsilon, \mathcal{F}, Q)$ is called ε-**entropy with bracketting** of F.*

One has to note that, *a priori* we only impose that the bounding functions belong to $L_p(Q)$ and not the entire set $\mathcal{F}$, but when the entropy with bracketing is finite, every function $f \in \mathcal{F}$ is at L_p-distance bounded by ε from an element f_i^L which belongs to $L_p(Q)$. Hence this impose that $\mathcal{F} \subset L_p(Q)$.

Exercice 17. *If $\mathbb{D} = \mathbb{R}$ and $A = \{x \in \mathbb{R} : |x| \leq k\}$ and $d = |\cdot|$ show that $\mathcal{N}(\varepsilon, A, d) \leq \lceil k/\varepsilon \rceil$.*

One has the following ordering between the different entropies.

Proposition 19. *For all $1 \leq p < \infty$ and $\forall \varepsilon > 0$,*

$$H_p(\varepsilon, \mathcal{F}, Q) \leq H_{p,B}(\varepsilon, \mathcal{F}, Q).$$

If Q is a measure of probability,

$$H_{p,B}(\varepsilon, \mathcal{F}, Q) \leq H_\infty\left(\frac{\varepsilon}{2}, \mathcal{F}\right)$$

If $A \subset \mathbb{D}$ and d, d' are two pseudo-distances on $\mathbb{D}$ such that $\forall x, y \in \mathbb{D}$, $d(x, y) \leq d'(x, y)$ then

$$H(\varepsilon, A, d) \leq H(\varepsilon, A, d').$$

One could have added, in the previous Proposition, the fact that if two metric spaces $(\mathbb{D}, d)$ and $(\mathbb{D}', d')$ are isometric, then there is a correspondence between the covering of $\mathbb{D}$ and the ones of $\mathbb{D}'$. We will use this fact without proof in the following examples.

Proof. Left as an exercice □

8.3.2 ε-nets

An ε**-net** of a set A is a finite family $(c_j)_{j=1,\dots,N}$ of elements of A such that

- For any $i \neq j$, $\|c_i - c_j\| \geq \varepsilon$,
- The set $\{c_1, \dots, c_N\}$ is maximal with respect to the inclusion order.

It is direct to see that there is a link between the covering number and the existence of an ε-net for a set A. This is formalized in the following result.

Proposition 20. *A ε-net $(c_i)_{i=1,\dots,N}$ of a set A forms the centers set Centers$(\mathcal{C})$ of a covering $\mathcal{C}$ of A.*

Proof. Let $(c_i)_{j=1,\dots,N}$ be a ε-net of A. The collection of the balls of radius ε centered at the c_j form a covering. Indeed, if it was not the case, we would be able to find a point $x \in A$ that do not belong to one of the balls $B_d(c_j, \varepsilon)$. That would mean that $\{c_1, \dots, c_N\} \cup \{x\}$ is also an ε-net of A which contradicts the maximality of the initial ε-net $\{c_1, \dots, c_N\}$. □

Lemma 9. *If $A = B_d(0, R) \subset \mathbb{R}^d$ endowed with the Euclidean distance d, then the covering number is such that*

$$\mathcal{N}(\varepsilon, A, d) \leq \left(\frac{2R + \varepsilon}{\varepsilon}\right)^d.$$

Proof. Let $(c_i)_{j=1,\dots,N}$ be a ε-net of the ball $B_d(R)$. By Proposition 20, it also forms a covering of $B_d(R)$ then we have $\mathcal{N}(\varepsilon, A, d) \leq N$. It is also true that

$$\bigcup_{j=1}^N B_d\left(c_j, \frac{\varepsilon}{2}\right) \subset B_d\left(R + \frac{\varepsilon}{2}\right).$$

The intersection of two balls $B_d\left(c_j, \frac{\varepsilon}{2}\right)$ is empty or reduced to a singleton. Hence one can compare the two Lebesgues measures of the previous sets to get

$$\sum_{j=1}^N \mu_d \left(\frac{\varepsilon}{2}\right)^d \leq \mu_d \left(R + \frac{\varepsilon}{2}\right)^d$$

where $\mu_d = 2\pi^{d/2} d^{-1} \Gamma(d/2)^{-1}$ is the volume of the unit ball in $\mathbb{R}^d$. Rearranging the last inequality gives the result. □

8.3.3 Examples

Example 5. *Let $\phi_1, \dots, \phi_d \in L_2(Q)$ fixed functions of unit norm and let*

$$\mathcal{F} = \left\{ f = \sum_{k=1}^{d} \theta_k \phi_k : \ \theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d, \|f\|_{2,Q} \leq R \right\}.$$

Then one has that for all $\varepsilon > 0$,

$$H_2(\varepsilon, \mathcal{F}, Q) \leq d_Q \log\left(\frac{2R+\varepsilon}{\varepsilon}\right)$$

where d_Q is the rank of the matrix $\Sigma_Q = \int \phi\phi^T dQ$ with the notation $\phi = (\phi_1, \dots, \phi_d)$. Indeed, one can see that there is a bijection that preserves the scalar product between $\mathcal{F}$ and the set of $\mathbb{R}^d$ given by

$$\left\{ u = \sum_{k=1}^{d} \theta_k e_k : \ \theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d, \|u\| \leq R \right\}$$

where the vectors e_k are such that $\forall i, j$, the scalar product is given by $e_i \cdot e_j = \int \phi_i \phi_j dQ$. Of course, if $(e_k)_k$ forms a orthonormal basis then the result holds with $d_Q = d$ by the use of Lemma 9. Otherwise, one can use the orthonormalization of Gram-Schmith to form an orthonormal basis $e'_1, \dots, e'_{d_Q}$ where d_Q is given by the rank of the Gram matrix $G = (e_i \cdot e_j)_{i,j}$. But since $G = \Sigma_Q$, one has the result.

Example 6. *Let*

$$\mathcal{F} = \{f : X \to [0,1] \text{ non-decreasing}\}$$

where X is a finite subset of $\mathbb{R}$. Then one has $H_\infty(\varepsilon, \mathcal{F}) \leq \varepsilon^{-1} \log(n + \varepsilon^{-1})$ where $n = |X|$. To see that, define $x_1 \leq \dots, \leq x_n$ the elements of X. We define, for all $f \in \mathcal{F}$,

$$M_i^f = \left\lfloor \frac{f(x_i)}{\varepsilon} \right\rfloor, \quad \forall i = 1, \dots, n.$$

Let $\tilde{f}(x_i) = \varepsilon M_i^f$, then $\|f - \tilde{f}\|_\infty \leq \varepsilon$. Also, the set of discretized functions $\tilde{\mathcal{F}} = \{\tilde{f} : \ f \in \mathcal{F}\}$ is finite since $1 \leq M_1^f \leq \dots \leq M_n^f \leq \lfloor \varepsilon^{-1} \rfloor$ are natural numbers. Exact computations give that

$$|\tilde{\mathcal{F}}| = \binom{n + \lfloor \varepsilon^{-1} \rfloor}{\lfloor \varepsilon^{-1} \rfloor} \leq (n + \lfloor \varepsilon^{-1} \rfloor)^{\lfloor \varepsilon^{-1} \rfloor}.$$

Since $\tilde{\mathcal{F}}$ induces a covering of $\mathcal{F}$, we get an upper bound of the covering number that gives the result.

Remark 2. *A famous result by Birman and Solomjak finally gives that $H_{1,B}(\varepsilon, \mathcal{F}, Q) \leq A\varepsilon^{-1}$ (see Chapter 9)*

Example 7. *Let*

$$\mathcal{F} = \{f : [0,1] \to [0,1] \text{ such that } |f'| \leq 1\}$$

then there exists a constant $A > 0$ such that

$$H_\infty(\varepsilon, \mathcal{F}) \leq \frac{A}{\varepsilon}, \quad \forall \varepsilon > 0.$$

To justify this, let $0 = a_0 <, \dots, a_N = 1$ such that $a_k = k\varepsilon$ for $k = 0, \dots, N-1$. Let $B_k = (a_{k-1}, a_k]$ and

$$\tilde{f} = \sum_{k=1}^{N} \varepsilon \left\lfloor \frac{f(a_k)}{\varepsilon} \right\rfloor \mathbb{1}_{B_k}.$$

We have that $\|f - \tilde{f}\| \leq 2\varepsilon$, by construction and the values of $\tilde{f}$ are the εM where M is an integer. Moreover,

$$|\tilde{f}(a_k) - \tilde{f}(a_{k-1})| \leq |\tilde{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \tilde{f}(a_{k-1})| \leq 3\varepsilon$$

To define the value of $\tilde{f}(a_0)$, we have $\lfloor \varepsilon^{-1} \rfloor + 1$ possibilities. Then for the choice of $\tilde{f}(a_1)$, the previous inequality only leave 7 possibilities. This is also 7 possibilities for $\tilde{f}(a_2)$ and so on. Finally, there is no more than $(\lfloor \varepsilon^{-1} \rfloor + 1)7^{\lfloor \varepsilon^{-1} \rfloor}$ such functions $\tilde{f}$. Then

$$H_\infty(2\varepsilon, \mathcal{F}) \leq \frac{1}{\varepsilon} \log 7 + \log(\frac{1}{\varepsilon} + 1) \leq \frac{A}{\varepsilon},$$

for A a universal constant.

8.4 A first result under entropy with bracketing

In the following, we will say that an empirical process $(\frac{1}{n}\sum_{i=1}^n f(X_i))_{f\in\mathcal{F}}$ is **P-Glivenko-Cantelli** when

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\left[f\right]\right| \xrightarrow{a.s.} 0.$$

This notion corresponds to a LLN that holds uniformly on the entire class $\mathcal{F}$. We now expose the simplest theorem using the finiteness of the entropy with bracketing for proving the uniform first order convergence of the empirical process. The following result is inspired by the proof of the classical Glivenko-Cantelli Theorem 5.

Theorem 10. *Let $\mathcal{F}$ be a class of functions. We assume that $H_{1,B}(\varepsilon,\mathcal{F},P)<\infty$, for all $\varepsilon>0$, then $\mathcal{F}$ is a class P-Glivenko-Cantelli.*

Proof. Let $\varepsilon>0$. By assumption, $N:=\mathcal{N}(\varepsilon,\mathcal{F},P)<\infty$ and then there exists a finite class $\{f_i^L,f_i^R\}_{i=1}^N$ such that $\|f_i^L-f_i^R\|\leq\varepsilon$ and $\forall f\in\mathcal{F}$, $\exists i$ such that $f_i^L\leq f\leq f_i^R$. Then

$$\begin{aligned}\int fd(P_n-P) &= \int fdP_n - \int fdP \leq \int f_i^R dP_n - \int fdP\\ &= \int f_i^R d(P_n-P) + \int (f_i^R - f)dP\\ &\leq \int f_i^R d(P_n-P) + \varepsilon.\end{aligned}$$

Similarly, we have that $\int fd(P_n-P)\geq\int f_i^L d(P_n-P)-\varepsilon$. Since $\{f_i^L,f_i^R\}_{i=1}^N$ is a finite set, a direct use of the classical LLN gives that

$$\max_{i=1,\dots,N}\left|\int f_i^L d(P_n-P)\right| \xrightarrow{a.s.} 0$$
$$\max_{i=1,\dots,N}\left|\int f_i^R d(P_n-P)\right| \xrightarrow{a.s.} 0.$$

Then, with probability 1, for n sufficiently large, one has that

$$\sup_{f\in\mathcal{F}}\left|\int fd(P_n-P)\right|\leq 2\varepsilon$$

and the result is proved. □

In fact, the finiteness of the entropy with bracketing has a second consequence that we expose in the following lemma that deals with the enveloppe of the class $\mathcal{F}$. The function

$$F=\sup_{f\in\mathcal{F}}|f|$$

is called **enveloppe** of the class $\mathcal{F}$.

Lemma 10. *Assume that $H_{1,B}(\varepsilon,\mathcal{F},P)<\infty$ for all $\varepsilon>0$. Then $F\in L_1(P)$.*

Proof. For every $\varepsilon>0$, $H_{1,B}(\varepsilon,\mathcal{F},P)$ is finite so is $H_1(\varepsilon,\mathcal{F},P)$ by Proposition 19. As a consequence, $(\mathcal{F},\|\cdot\|_{1,P})$ is totally bounded and then $\overline{\mathcal{F}}$ is pre-compact. It is also immediate to see that every function $f\in\mathcal{F}$ belongs to $L_1(P)$ since it is at L_1-distance bounded by ε of a function in $L_1(P)$. Since the space $L_p(Q)$ is complete we have that $\overline{\mathcal{F}}$ is also complete. But since a pre-compact set which is also complete is compact (this is actually an equivalence), we have that $\overline{\mathcal{F}}$ is compact. Moreover, $f\mapsto\|f\|_{1,P}$ is a continuous function, it is a bounded function (that also attains its bounds). Then, there exists $R>0$ such that

$$\sup_{f\in\mathcal{F}}\|f\|_{1,P}\leq R.$$

Now, fix $\varepsilon>0$, so that for any function $f\in\mathcal{F}$, we have that $f_i^L\leq f\leq f_i^R$ and then

$$|f|\leq|f_i^L|+|f_i^R-f_i^L|\leq\sum_{i=1}^N|f_i^L|+|f_i^R-f_i^L|$$

where $N = \exp(H_{1,B}(\varepsilon, \mathcal{F}, P))$. Then

$$\|F\|_{1,P} \leq \sum_{i=1}^{N} \|f_i^L\|_{1,P} + \|f_i^R - f_i^L\|_{1,P} \leq N(R + 2\varepsilon).$$

This insures that $F \in L_1(P)$. □

This last result gives an indication on the minimal assumptions that one would impose to have the uniform LLN. Indeed, one has the fact that the last result is actually necessary (under the extra condition that the set $\mathcal{F}$ is bounded in $L_1(P)$). This last hypothesis is, of course, necessary since one can think of the P-Glivenko-Cantelli class of the constant functions that do not have an integrable enveloppe. In the other theorem that we will present (Theorem 11), this necessary condition will be assumed.

Proposition 21. *If the class $\mathcal{F}$ is P-Glivenko-Cantelli and bounded in $L_1(P)$, then $F \in L_1(P)$.*

Proof. Since $\mathcal{F}$ is P-Glivenko-Cantelli, $\sup_{f\in\mathcal{F}} |P_n f - Pf| \xrightarrow{a.s.} 0$. But

$$\frac{1}{n} \sup_{f\in\mathcal{F}} |f(X_n) - Pf| \leq \sup_{f\in\mathcal{F}} |P_n f - Pf| + \left(1 - \frac{1}{n}\right) \sup_{f\in\mathcal{F}} |P_{n-1} f - Pf|$$

then $n^{-1} \sup_{f\in\mathcal{F}} |f(X_n) - Pf| \xrightarrow{a.s.} 0$ which implies that $\mathbb{P}\left(\sup_{f\in\mathcal{F}} |f(X_n) - Pf| \geq n, \text{i.o.}\right) = 0$. By Borel-Cantelli Lemma 33, one has that $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |f(X_n) - Pf|\right] \leq \sum_n \mathbb{P}\left(\sup_{f\in\mathcal{F}} |f(X_n) - Pf| \geq n\right) < \infty$. The random variables X_i are i.i.d. so that we actually proved that $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |f(X) - Pf|\right] < \infty$. Since $\mathcal{F}$ is bounded in $L_1(P)$, we have that

$$\mathbb{E}[F] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |f(X) - Pf|\right] + \sup_{f\in\mathcal{F}} |Pf| < +\infty.$$

□

8.5 A second result under empirical entropy control

The objective of this section is to prove the following theorem.

Theorem 11. *If the enveloppe F of $\mathcal{F}$ is in $L_1(P)$ and if*

$$\frac{1}{n} H_1(\varepsilon, \mathcal{F}, P_n) \xrightarrow{\mathbb{P}} 0, \quad \forall \varepsilon > 0,$$

then $\mathcal{F}$ is P-Glivenko-Cantelli.

This results is much weaker than Theorem 10 in two perspectives. First, the condition holds on a notion of entropy that is smaller since, by Proposition 19 the H_1 entropy is bounded by the entropy with bracketing $H_{1,B}$. Secondly, the order of magnitude is bigger ($o_P(n)$ against the $O(1)$ for Theorem 10) which allows a little more freedom in the research of upper bounds for the entropies. Nonetheless, the price to pay is to deal with an entropy that is now a random variable.

Proof. See Section [XXX] □

8.5.1 Symmetrization

We will use the following lemma in the proof of Theorem 11. More results of this flavor can be found in the excellent [4]. This kind of results link the theory of empirical processes to the theory of Rademacher chaos where another notion of complexity for sets is defined. This complexity is the so called Rademacher complexity. [Develop this point]

Lemma 11. *Let $X_1, \ldots, X_n$ be independent random processes $X_i = (X_{i,s})_{s\in\mathcal{T}}$ assumed centered (i.e. $\forall i$, $\forall s$, $\mathbb{E}[X_{i,s}] = 0$). Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables and independent from $X_1, \ldots, X_n$, then*

$$\frac{1}{2}\mathbb{E}\left[\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^{n} \varepsilon_i X_{i,s}\right|\right] \underset{(2)}{\leq} \mathbb{E}\left[\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^{n} X_{i,s}\right|\right] \underset{(1)}{\leq} 2\mathbb{E}\left[\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^{n} \varepsilon_i X_{i,s}\right|\right]$$

and

$$\mathbb{E}\left[\sup_{s\in\mathcal{T}} \sum_{i=1}^{n} X_{i,s}\right] \underset{(3)}{\leq} 2\mathbb{E}\left[\sup_{s\in\mathcal{T}} \sum_{i=1}^{n} \varepsilon_i X_{i,s}\right]$$

Proof. We begin with the proof of (1). Since the processes X_i are centered, it holds that

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^n X_{i,s}\right|\right] &= \mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^n X_{i,s} - \mathbb{E}\left[X'_{i,s}\right]\right|\right] \\
&= \mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\mathbb{E}\left[\sum_{i=1}^n X_{i,s} - X'_{i,s}\middle| X'_1,\dots,X'_n\right]\right|\right] \\
&\le \mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^n (X_{i,s} - X'_{i,s})\right|\right] \quad \text{(by Jensen's inequality used twice)} \\
&= \mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i(X_{i,s} - X'_{i,s})\right|\right] \quad \text{(by symmetry of } X_{i,s}-X'_{i,s} \text{ in distribution)} \\
&\le 2\mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i X_{i,s}\right|\right] \quad \text{(by triangular inequality)}
\end{aligned}$$

where $X'_{i,s}$ is an independent copy of the random variable $X_{i,s}$. The inequalities (2) and (3) can be proved in a very similar manner. □

The symmetrization that we used in Lemma 11 is a general idea that can also be used to prove that the concentration of the empirical process is of the same order of its symmetrized version. More formally, we have the following result.

Lemma 12. *Assume that for any function $f \in \mathcal{F}$ and a $\delta > 0$,*

$$\mathbb{P}\left(\left|\int f d(P_n - P)\right| > \frac{\delta}{2}\right) \le \frac{1}{2}.$$

Then, it holds that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f d(P_n - P)\right| > \delta\right) \le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f d(P_n - P'_n)\right| > \frac{\delta}{2}\right)$$

where P'_n is the empirical measure defined on $(X'_1,\dots,X'_n)$ which is a independent copy of $(X_1,\dots,X_n)$.

Proof. Denote by $\mathbf{X}$ the vector $(X_1,\dots,X_n)$ and by $A_f = \{\mathbf{X} :\ |\int f d(P_n - P)| > \delta\}$. We also define $A = \bigcup_{f\in\mathcal{F}} A_f$. By definition of A, if $\mathbf{X} \in A$ means that there exists $f^* = f^*_{\mathbf{X}} \in \mathcal{F}$ such that $\mathbf{X} \in A_{f^*}$. As a function dependent of $\mathbf{X}$, f^* is then a random function in $\mathcal{F}$. By independence of P_n and P'_n,

$$\begin{aligned}
\mathbb{P}\left(A_{f^*} \text{ and } \left|\int f^* d(P'_n - P)\right| \le \frac{\delta}{2}\right) &= \mathbb{E}_{\mathbf{X}}\left[\mathbb{P}_{\mathbf{X}'}\left(\left|\int f^* d(P'_n - P)\right| \le \frac{\delta}{2}\right)\mathbb{1}_{A_{f^*}}\right] \\
&> \frac{1}{2}\mathbb{P}\left(A_{f^*}\right) = \frac{1}{2}\mathbb{P}\left(\left|\int f^* d(P_n - P)\right| > \delta\right).
\end{aligned}$$

Using this inequality, we find that

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f d(P_n - P)\right| > \delta\right) &= \mathbb{P}\left(\mathbf{X} \in \bigcup_{f\in\mathcal{F}} A_f\right) \\
&\le \mathbb{P}\left(\mathbf{X} \in A_{f^*}\right) = \mathbb{P}\left(\left|\int f^* d(P_n - P)\right| > \delta\right) \\
&\le 2\mathbb{P}\left(\mathbf{X} \in A_{f^*} \text{ and } \left|\int f^* d(P'_n - P)\right| \le \frac{\delta}{2}\right) \\
&= 2\mathbb{P}\left(\left|\int f^* d(P_n - P)\right| > \delta \text{ and } \left|\int f^* d(P'_n - P)\right| \le \frac{\delta}{2}\right) \\
&\le 2\mathbb{P}\left(\left|\int f^* d(P_n - P'_n)\right| > \frac{\delta}{2}\right) \\
&\le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f d(P_n - P'_n)\right| > \frac{\delta}{2}\right)
\end{aligned}$$

□

To get a result of the form of Lemma 11, one can apply the Rademacher random variables trick and get the following result.

Corollary 5. *Let $\varepsilon_1, \ldots, \varepsilon_n$ be Rademacher random variables independent from the X_i. Then, under the hypothesis of Lemma 12,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int fd(P_n - P)\right| > \delta\right) \le 4\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| > \frac{\delta}{4}\right).$$

Proof. The proof used the same ideas as the proof of Lemma 11 and is left as an exercice. □

[STATE A RESULT ABOUT THE LINK WITH RADEMACHER COMPLEXITY]

8.5.2 Dudley entropy integral

In this section we derive a result that is the starting point of a general theory known under the name of chaining technique. This idea was first introduced by Kolmogorov [7]. The idea is to reduce a supremum over an infinite class to a supremum over increments of a process where each increment can only take a finite number of values. The original idea comes from Dudley [6] and further studied and extended by Talagrand (see [12]).

Lemma 13. *Let $X_1, \ldots, X_N$ be sub-gaussian random variables $\mathcal{G}(v)$ (i.e. $\forall\lambda > 0$, $\mathbb{E}\left[e^{\lambda X_i}\right] \le e^{\lambda^2 v/2}$). Then*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \le \sqrt{2v\log N}.$$

Proof. By Jensen's inequality, $\forall\lambda > 0$,

$$\begin{aligned}
\exp\left(\lambda\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right]\right) &\le \mathbb{E}\left[\exp\left(\lambda\max_{i=1,\ldots,N} X_i\right)\right] \\
&= \mathbb{E}\left[\max_{i=1,\ldots,N}\exp\left(\lambda X_i\right)\right] \\
&\le \sum_{i=1}^N \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \le N\exp\left(\frac{\lambda^2 v}{2}\right)
\end{aligned}$$

Taking the logarithm, we get that for all $\lambda > 0$,

$$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \le \frac{\log(N)}{\lambda} + \frac{\lambda v}{2}.$$

Since the left hand side does not depend on λ, one can minimize in λ the right hand side. Hence, taking $\lambda = \sqrt{2\log(N)/v}$, we get the result. □

Theorem 12 (Dudley entropy integral)**.** *Let $(\mathcal{T}, d)$ be a metric space and let $(X_t)_{t\in\mathcal{T}}$ be a process indexed by $\mathcal{T}$ such that, for all $t, t' \in \mathcal{T}$ and all $\lambda > 0$,*

$$\log\mathbb{E}\left[\exp\lambda(X_t - X_{t'})\right] \le \frac{\lambda^2 d^2(t,t')}{2}.$$

Then, for every $t_0 \in \mathcal{T}$,

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}|X_t - X_{t_0}|\right] \le 12\int_0^{\delta/2}\sqrt{H(\varepsilon,\mathcal{T},d)}d\varepsilon \tag{8.3}$$

where $\delta = \sup_{t\in\mathcal{T}} d(t,t_0)$. In particular, for $D = diam(\mathcal{T})$,

$$\mathbb{E}\left[\sup_{t,s\in\mathcal{T}}(X_t - X_s)\right] \le 24\int_0^{D/2}\sqrt{H(\varepsilon,\mathcal{T},d)}d\varepsilon. \tag{8.4}$$

Proof. We assume that the metric entropy is finite for any $\varepsilon > 0$ otherwise the bound is trivial. Actually, one can only take $H(\varepsilon,\mathcal{T},d)$ finite a.s. but the following is trivially adaptable to this general case. We start by assuming that $\mathcal{T}$ is finite. For any $j \in \mathbb{N}$, we define $\delta_j = \delta 2^{-j}$. For any $j \in \mathbb{N}$, $N_j := \mathcal{N}(\delta_j,\mathcal{T},d)$ is finite and then there exists a finite covering $\bigcup_{i=1}^{N_j} B_d(x_i,\delta_j)$ of $\mathcal{T}$. Let $\mathcal{T}_j$ be the finite ensemble of the centers of that covering. For every $j \in \mathbb{N}$, we define a function $\Pi_j : \mathcal{T} \to \mathcal{T}_j$ that associated any $t \in \mathcal{T}$ to a point in $\mathcal{T}_j$ such that $d(t,\Pi_j(t)) \le \delta_j$. There may be more than one possibility for $\Pi_j(t)$. When it is the case, one may choose any of the candidates arbitrarily. We finally define $\mathcal{T}_0 = \{t_0\}$

and $\Pi_0(t) = t_0$.
Step 1: We have that,

$$X_t = X_{t_0} + \sum_{j=0}^{\infty} X_{\Pi_{j+1}(t)} - X_{\Pi_j(t)}.$$

Indeed, since the set $\mathcal{T}$ is finite, the infinite sum is actually finite and for j large enough $X_{\Pi_j(t)} = X_t$.
Step 2: Then one has that

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_t - X_{t_0}|\right] \leq \sum_{j=0}^{\infty} \mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_{\Pi_{j+1}(t)} - X_{\Pi_j(t)}|\right]$$

Furthermore, $|\{(\Pi_j(t), \Pi_{j+1}(t)) : t \in \mathcal{T}\}| \leq \exp(2H(\delta_{j+1}, \mathcal{T}, d))$ and the triangular inequality of d gives

$$d(\Pi_j(t), \Pi_{j+1}(t)) \leq d(\Pi_j(t), t) + d(t, \Pi_{j+1}(t)) = \delta_j + \delta_{j+1} = 3\delta_{j+1}.$$

Then the variables $X_{\Pi_{j+1}(t)} - X_{\Pi_j(t)}$ are sub-Gaussian of constant $9\delta_{j+1}^2$ so that one can use Proposition 13 to get

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_{\Pi_{j+1}(t)} - X_{\Pi_j(t)}|\right] \leq \sqrt{2 \times 9\delta_{j+1}^2 \times 2H(\delta_{j+1}, \mathcal{T}, d)} = 6\delta_{j+1}\sqrt{H(\delta_{j+1}, \mathcal{T}, d)}.$$

Then summing these inequalities we get

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_t - X_{t_0}|\right] \leq \sum_{j=1}^{\infty} 6\delta_j \sqrt{H(\delta_j, \mathcal{T}, d)} = 12\sum_{j=1}^{\infty}(\delta_j - \delta_{j+1})\sqrt{H(\delta_j, \mathcal{T}, d)} \leq 12\int_0^{\delta/2} \sqrt{H(\varepsilon, \mathcal{T}, d)}d\varepsilon$$

where, in the last step we used the classical comparison of Riemann sums and integrals on the non-increasing function $\delta \mapsto H(\delta, \mathcal{T}, d)$. As a by product of the result, we obtained that for any $\varepsilon > 0$, there exists $\eta > 0$ such that for any finite and thus countable subset S of $\mathcal{T}$,

$$\mathbb{E}\left[\sup_{\substack{s,t\in S\\ d(s,t)<\eta}} |X_s - X_t|\right] \leq \varepsilon. \tag{8.5}$$

Step 3: In the general case, by the assumption on the finiteness of the integral, $(\mathcal{T}, d)$ is totally bounded and then separable. Then, there exists a countable set S dense in $\mathcal{T}$. Let us take $\tilde{X}_t = X_t$ for any $t \in S$ and $\tilde{X}_t = \lim X_s$ where the limit is in the L_1 sense by the help of Equation (8.5). Then $(\tilde{X}_t)_{t\in\mathcal{T}}$ is modification (see Definition 14 for a concrete definition) of $(X_t)_{t\in\mathcal{T}}$ that has a.s. continuous paths. By continuity of the paths, Equations (8.3) and (8.4) are satisfied for the process $(\tilde{X}_t)_{t\in\mathcal{T}}$ where the sup is taken on the complete $\mathcal{T}$. We finish the proof by saying that, by construction,

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_t - X_{t_0}|\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}} |\tilde{X}_t - \tilde{X}_{t_0}|\right]$$

where the left hand side expectation has to be taken as one of the generalized expectations of (8.1) or (8.2). □

As a by product of the proof of Theorem 12, we get that one can construct a continuous version of the process $(X_t)_t$ when the entropy integral is finite. In fact, Theorem 12 can be generalized for other classes of random variables and then have as a consequence the famous Kolmogorov continuity theorem. This aspect is briefly treated in Section [XXX].

Remark 3. *It is a theorem that makes the link between a discrete and finite case to a continuous and infinite case. As the vigilant reader may have noticed, the only use of the distance property is made through the triangular inequality. As a direct consequence, the same theorem is true for spaces $(\mathcal{T}, d)$ where d is only a pseudo metric. Of course, the definition of entropy does not change under this alternative setting.*

8.5.3 Sudakov Minoration

Proposition 22. *Let $X_1, \dots, X_n$ ne i.i.d. standard Gaussian random variables. Then, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}\left[\max_{i=1,\dots,N} X_i\right] \geq C\sqrt{\log(N)} - 2$$

Moreover, it holds that

$$\frac{\mathbb{E}\left[\max_{i=1,\dots,N} X_i\right]}{\sqrt{2\log(N)}} \underset{N\to\infty}{\longrightarrow} 1.$$

Proof. We start with the absolute value of standard Gaussian random variables. For any $\delta > 0$, one has that

$$\mathbb{E}\left[\max_{i=1,\dots,N}|X_i|\right] = \int_0^\infty \mathbb{P}\left(\max_{i=1,\dots,N}|X_i| > t\right)dt \geq \int_0^\delta \mathbb{P}\left(\max_{i=1,\dots,N}|X_i| > \delta\right)dt = \delta[1-(1-\mathbb{P}(|X_1|>\delta))^N].$$

But $\mathbb{P}(|X_1|>\delta) = \sqrt{2/\pi}\int_\delta^\infty e^{-t^2/2}dt = 1-\sqrt{2/\pi}\int_0^\delta e^{-t^2/2}dt$. But

$$\int_0^\delta e^{-t^2/2}dt = \sqrt{\int_0^\delta\int_0^\delta e^{-t_1^2/2-t_2^2/2}dt_1dt_2} \leq \sqrt{\int_0^{\pi/2}\int_0^{\sqrt{2}\delta}\rho e^{-\rho^2/2}d\rho d\theta} = \sqrt{\frac{\pi}{2}(e^{-\delta^2}-1)},$$

so that $\mathbb{P}(|X_1|>\delta) \geq e^{-\delta^2/2}$. Now take $\delta = \sqrt{2\log(N)}$, then $\mathbb{P}(|X_1|>\delta) \geq 1/N$ and

$$\mathbb{E}\left[\max_{i=1,\dots,N}|X_i|\right] \geq \delta\left[1-(1-\frac{1}{N})^N\right] \geq (1-e^{-1})\sqrt{2\log(N)}.$$

But since $\mathbb{E}\left[\max_{i=1,\dots,N}|X_i|\right] \leq \mathbb{E}\left[\max_{i=1,\dots,N}X_i\right] + 2\mathbb{E}\left[|X_1|\right]$, one has that

$$\mathbb{E}\left[\max_{i=1,\dots,N}X_i\right] \geq (1-e^{-1})\sqrt{2\log(N)}-2.$$

The constant e^{-1} can be reduced by taking for example $\delta = \sqrt{2\log(pN)}$ the bound becomes $(1-e^{-p})\sqrt{2\log(pN)}-2$ and we get the second result by taking $p\to\infty$. □

8.6 Proof of Theorem 11

In order to prove Theorem 11, we give three successive lemmas that use Dudley bound along with symetrization.

Lemma 14. *Let $X_1,\dots,X_n$ be i.i.d. random variables and let $\varepsilon_1,\dots,\varepsilon_n$ be i.i.d. Rademacher Rad$(1/2)$ random variables independent from the variables $X_1,\dots,X_n$. Then,*

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] \leq \frac{\sup_{f\in\mathcal{F}}\|f\|_{2,P_n}}{\sqrt{n}} + 12\int_0^{D_n}\sqrt{\frac{H(\varepsilon,\mathcal{F},\|\cdot\|_{2,P_n})}{n}}d\varepsilon$$

where $D_n = \sup_{f\in\mathcal{F}}\|f\|_{n,\infty}$ for $\|f\|_{n,\infty} = \max_{i=1,\dots,n}|f(X_i)|$ and $\|f\|_{2,P_n}^2 = n^{-1}\sum_{i=1}^n f(X_i)^2$. The notation $\mathbb{E}_\varepsilon$ holds for the expectation operator on the random variables ε_i at X_i fixed.

Proof. In order to use Dudley's bound, one has to verify that the increments of the process $(\sum_{i=1}^n\varepsilon_i f(X_i))_f$ are sub-Gaussian. For any two functions f and f', the random variable $\sum_{i=1}^n\varepsilon_i(f(X_i)-f'(X_i))$ can be seen as a Rademacher chaos of order 1 and hence is a sum of independent and bounded random variables. This fact allows us to use Hoeffding's inequality given in Theorem 9. This gives that the increment is sub-Gaussian of constant $n^{-2}\sum_{i=1}^n(f(X_i)-f'(X_i))^2 = (\frac{1}{\sqrt{n}}\|f-f'\|_{2,P_n})^2$. We use Dudley entropy bound on the set $\mathcal{F}$, then for a $f_0\in\mathcal{F}$,

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)-f_0(X_i)\right|\right] \leq 12\int_0^{\delta_n/2}\sqrt{H(\varepsilon,\mathcal{F},n^{-1/2}\|\cdot\|_{2,P_n})}d\varepsilon = 12\int_0^{\delta_n/2}\sqrt{H(\sqrt{n}\varepsilon,\mathcal{F},\|\cdot\|_{2,P_n})}d\varepsilon$$

where $\delta_n = n^{-1/2}\sup_f\|f-f_0\|_{2,P_n}$. But using Hoeffding inequality on $\sum_{i=1}^n\varepsilon_i f_0(X_i)$ we get that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f_0(X_i)\right|\right] = \int_0^\infty\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f_0(X_i)\right|\geq t\right)dt \leq \int_0^\infty 2e^{(-2\frac{n^2t^2}{\sum f_0(X_i)^2})}dt = \frac{\sqrt{2\pi}\|f_0\|_{2,P_n}}{4\sqrt{n}} \leq \frac{\sup\|f\|_{2,P_n}}{\sqrt{n}}$$

and by a change of variables $\varepsilon' = \sqrt{n}\varepsilon$, we get that

$$\int_0^{\delta_n/2}\sqrt{H(\sqrt{n}\varepsilon,\mathcal{F},\|\cdot\|_{2,P_n})}d\varepsilon = \int_0^{\sqrt{n}\delta_n/2}\sqrt{\frac{H(\varepsilon',\mathcal{F},\|\cdot\|_{2,P_n})}{n}}d\varepsilon' \leq \int_0^{D_n}\sqrt{\frac{H(\varepsilon,\mathcal{F},\|\cdot\|_{2,P_n})}{n}}d\varepsilon$$

□

Of course, one has that $H(\varepsilon,\mathcal{F},\|.\|_{2,P_n}) = H_2(\varepsilon,\mathcal{F},P_n)$.

Lemma 15. *Let $R > 0$ and assume that $\sup_{f\in\mathcal{F}} \|f\|_\infty < R$. If*

$$\frac{1}{n} H_2(\varepsilon, \mathcal{F}, P_n) \xrightarrow{\mathbb{P}} 0, \quad \forall \varepsilon > 0$$

then $\mathcal{F}$ is P-Glivenko-Cantelli.

Proof. Since the random variables are uniformly bounded, Remark 4 can be applied and then one only have to prove the weaker $\sup_f |P_n f - Pf| \xrightarrow{\mathbb{P}} 0$. So one can prove that the convergence occurs in L_1 to have the result. By the symmetrization argument prove in Lemma 11 and Lemma 14, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n f - Pf|\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] = \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]\right] \leq \frac{R}{\sqrt{n}} + 24\int_0^R \sqrt{\frac{H_2(\varepsilon, \mathcal{F}, P_n)}{n}} d\varepsilon.$$

The first term tend to 0, then one only has to prove that the integral tends to 0. To use Proposition 1 to prove that the random variables of interest are U.I., we notice that by brute force

$$\mathcal{N}_2(\varepsilon, \mathcal{F}, P_n) \leq \mathcal{N}_\infty(\varepsilon, \mathcal{F}) \leq \left(\frac{R}{\varepsilon}\right)^n,$$

which is an integrable function, so that,

$$Y_n := \int_0^R \sqrt{\frac{H_2(\varepsilon, \mathcal{F}, P_n)}{n}} d\varepsilon \leq \int_0^R \sqrt{\frac{n \log\left(\frac{R}{\varepsilon}\right)}{n}} d\varepsilon = \int_0^R \sqrt{\log(R/\varepsilon)} d\varepsilon < \infty.$$

Then, by the dominated convergence theorem (in its probabilistic version given by Corollary 8), $Y_n \xrightarrow{\mathbb{P}} 0$ and since $(Y_n)_n$ is U.I. (because bounded by Proposition 1) we also have that $Y_n \xrightarrow{\mathbb{L}_1} 0$ which concludes the proof. □

We are now ready to tackle the proof of our main theorem of this section.

Proof of Theorem 11. For a $R > 0$, we define the truncated set $\mathcal{F}_R = \{f\mathbb{1}_{F\leq R} : f \in \mathcal{F}\}$ where F is the enveloppe of the functions in $\mathcal{F}$. For two functions $f_1, f_2 \in \mathcal{F}$, the function $g_1 = f_1\mathbb{1}_{F\leq R}$ and $g_2 = f_2\mathbb{1}_{F\leq R}$ belong to $\mathcal{F}_R$. But then

$$\int (g_1 - g_2)^2 dP_n = \int_{F\leq R} (f_1 - f_2)^2 dP_n \leq 2R \int |f_1 - f_2| dP_n$$

which show that the assumption $n^{-1}H_1(\varepsilon, \mathcal{F}, P_n) \xrightarrow{\mathbb{P}} 0$ implies that $n^{-1}H_2(\varepsilon, \mathcal{F}_R, P_n) \xrightarrow{\mathbb{P}} 0$. Then, by Lemma 15, one has that the set of functions $\mathcal{F}_R$ is P-Glivenko-Cantelli. Now, by integrability of F, for any $\delta > 0$ there exists $R_0 > 0$ such that $\int_{F\geq R_0} F dP \leq \delta$. Since the trivial set $\{F\mathbb{1}_{F\geq R_0}\}$ and $\mathcal{F}_{R_0}$ are P-Glivenko-Cantelli, one have that for n large enough,

$$\sup_{f\in\mathcal{F}} \left|\int_{F\leq R_0} f d(P_n - P)\right| \leq \delta \quad \text{a.s.} \qquad \text{and} \qquad \int_{F\geq R_0} F dP_n \leq 2\delta \quad \text{a.s.}$$

Finally, we write

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left|\int f d(P_n - P)\right| &\leq \sup_{f\in\mathcal{F}} \left|\int_{F\leq R_0} f d(P_n - P)\right| + \sup_{f\in\mathcal{F}} \left|\int_{F\geq R_0} f d(P_n - P)\right| \\
&\leq \sup_{f\in\mathcal{F}} \left|\int_{F\leq R_0} f d(P_n - P)\right| + \int_{F\geq R_0} F dP_n + \int_{F\geq R_0} F dP \\
&\leq 4\delta \quad \text{a.s.}
\end{aligned}$$

which finishes the proof. □

Chapter 9

Birman and Solomjak theory

In this chapter, we derive the calculation leading to concrete calculations on the entropy of various sets of functions with enough regularity. This theory is taken from the seminal paper [1].

9.1 Notations and definitions

9.1.1 Functional space $W_p^\alpha(\Delta)$ and $V_\beta(\Delta)$

Let Q^m be the m-dimensional half-open unit cube in $\mathbb{R}^m$ (i.e. $0 \leq x_i < 1$, $i = 1, \ldots, m$). We denote by $k = (k_1, \ldots, k_m)$ a multi-index ($\forall i$, k_i is an non-negative integer), $x_k = \prod_{i=1}^m x_i^{k_i}$ and $|k| = \sum k_i$. We denote by D^k the corresponding diferencial operator given by

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \ldots \partial x_m^{k_m}}.$$

For a cube Δ with edges parallel to the coordinate axes, $p \geq 1$, $\alpha > 0$ we denote by $W_p^\alpha(\Delta)$ the Sobolev space endowed with its natural norm $\|\cdot\|_{W_p^\alpha(\Delta)}$. We recall that for $\theta = \alpha - \lfloor \alpha \rfloor$ and $u \in W_p^\alpha(\Delta)$,

$$\|u\|_{W_p^\alpha(\Delta)} = \|u\|_{L_p(\Delta)} + \|u\|_{L_p^\alpha(\Delta)}$$

where

$$\|u\|_{L_p^\alpha(\Delta)} = \sum_{|k|=\alpha} \int_\Delta |D^k u|^p dx$$

if α is an integer or

$$\|u\|_{L_p^\alpha(\Delta)} = \sum_{|k|=\alpha} \int_\Delta \int_\Delta \frac{|D^k u(x) - D^k u(y)|^p}{|x-y|^{p\theta+m}} dxdy$$

otherwise. The semi-norm $\|\cdot\|_{L_p^\alpha(\Delta)}$, has a homogeneity property with respect to linear transformation of the cube. For example, if one takes $\Delta = x_0 + hQ^m$, then

$$\|u\|_{L_p^\alpha(\Delta)} = h^{mp^{-1}-\alpha} \|u\|_{L_p^\alpha(Q^m)}.$$

In the one dimensional case (Δ is then an interval), we will use the notion of function of bounded β-variation denoted by $V_\beta(\Delta)$. Let $\beta \geq 1$. We say that $u \in V_\beta(\Delta)$, if

$$\|u\|_{V_\beta^0(\Delta)}^\beta = \sup \sum_{i=1}^n |u(x_i) - u(x_{i-1})|^\beta$$

is finite. The suprema is taken over all the possible finite sets of points $x_0 < x_1 < \cdots < x_n$ in the interval Δ. Of course, the set $V_\beta(\Delta)$ is a Banach space relatively to the norm

$$\|u\|_{V_\beta(\Delta)} = \|u\|_{V_\beta^0(\Delta)} + \sup_{x \in \Delta} |u(x)|.$$

9.1.2 Partitions Λ

In this section we consider partition of the cube Q^m where the elements are also m-dimensional cubes, generally denoted by Λ. We denote by $|\Lambda|$ the number of cubes in this partition and $\Lambda = \{\Delta_1, \ldots, \Delta_{|\Lambda|}\}$. A elementary extension of the partition Λ is a partition Λ' obtained by dividing some cubes in Λ into 2^m smaller cubes (by slicing in every dimension). The notation Λ_0 holds for the trivial partition.

Cube argument functions We define a non-negative function J on the half open cubes Δ that is semiadditive from below: For any partition of Δ into smaller cubes Δ_j,

$$\sum_j J(\Delta_j) \leq J(\Delta).$$

Let $|\Delta|$ be the Euclidean volume of the cube Δ and let $a > 0$. We define

$$g_a(J, \Delta) = |\Delta|^a J(\Delta)$$

and the following function of a partition Λ

$$G_a(J, \Lambda) = \max_{\Delta \in \Lambda} g_a(J, \Delta).$$

Slicing strategy One wants to track the minimal value of G_a given that the partitions considered have at most a certain number of elements. In other words, one is looking to

$$\begin{aligned} \textbf{Minimize} \quad & \Lambda \mapsto G_a(J, \Lambda) \\ \textbf{where} \quad & |\Lambda| \leq n. \end{aligned} \tag{9.1}$$

One employs a strategy of successive divisions. The first step is to divide Q^m into 2^m cubes and call Λ_1 the partition obtained. Then one partition again the cubes for which $g_a(J, \Delta)$ is such that

$$g_a(J, \Delta_j) \geq 2^{-ma} G_a(J, \Lambda_1)$$

into 2^m smaller cubes. This last step is then repeated. This constructs a sequence $T_a(J) = (\Lambda_i)_{i \in \mathbb{N}}$ of partitions such that Δ_{i+1} is an elementary extension of Δ_i.

9.1.3 Two elementary lemmas

Lemma 16. *Suppose that a cube Δ in Q^m is divided into cubes Δ_j for $j = 1, \dots, 2^m$. Then*

$$\max_j g_a(J, \Delta_j) \leq 2^{-ma} g_a(J, \Delta)$$

Proof. By the semiadditivity from below, we have that $\sum_j J(\Delta_j) \leq J(\Delta)$. But for all j, $|\Delta_j|^a = |\Delta|^a 2^{-ma}$ and then the maximum being upper bounded by the sum, we get the result. ☐

Lemma 17. *Let $s \in \mathbb{N}$ and let $x_j > 0$, $y_j > 0$ $(j = 1, \dots, s)$ be numbers such that*

$$\sum_j x_j \leq 1, \quad \sum_j y_j \leq 1, \quad x_j y_j^a \geq b \quad (j = 1, \dots, s).$$

for some $a > 0$ and $b > 0$. Then $b \leq s^{-(a+1)}$.

Proof. This is a classical optimization problem that one can tackle with Lagrange multipliers. Indeed, we look for $\max b$ satisfying the conditions of the lemma. Then one has to find the unique critical point of

$$b, (x_j)_j, (y_j)_j, (\lambda_j)_j, \alpha, \beta \mapsto b + \sum_j \lambda_j (x_j y_j^a - b) + \alpha(1 - \sum_j x_j) + \beta(1 - \sum_j y_j)$$

One has to verify the $3s + 3$ equations

$$\begin{cases} \sum_j x_j = 1 & (L_1) \\ \sum_j y_j = 1 & (L_2) \\ \sum_j \lambda_j = 1 & (L_3) \end{cases}, \quad \begin{cases} x_j y_j^a - b = 0 & (L_{1,j}) \\ \lambda_j y_j^a - \alpha = 0 & (L_{2,j}) \\ a \lambda_j x_j y_j^{a-1} - \beta = 0 & (L_{3,j}) \end{cases} \quad (j = 1, \dots, s).$$

For example from $(L_{1,j})$, $(L_{2,j})$ we get that $\lambda_j b = \alpha x_j$ and then $b = \alpha$ together with $\lambda_j = x_j$. In the same way, we get that $\lambda_j = y_j$. Consequently, the sequences $(x_j)_j$, $(y_j)_j$ and $(\lambda_j)_j$ are stationary so that $\forall j, x_j = y_j = \lambda_j = s^{-1}$. It gives $\max b = s^{-(a+1)}$ and the result follows. ☐

Chapter 10

M-estimation

The M-estimation (M for maximum) is a commonly used technique in statistics to define estimators of the "best" kind for a given problem. They are based on the minimization of some random criteria that measures the desired quality of the estimation.

10.1 Introduction and notations

Let $X_1, \dots, X_n, X$ be i.i.d. random variables taking values in a set $\mathcal{X}$ of common distribution P. Let $\mathcal{S}$ denote the set of parameters. In this chapter, $\mathcal{S}$ is assumed to be a subset of a metric set, so that it is possible to enroll $\mathcal{S}$ with a distance d. A random criteria is a function

$$\begin{aligned}\gamma_n \colon \mathcal{S} &\to \mathbb{R}_+^* \\ t &\mapsto \gamma_n(t) := \gamma_n(X_1, \dots, X_n, t)\end{aligned}$$

depending on the random variables $X_1, \dots, X_n$.

Settings and M-estimator Once given the criteria γ_n, one is interested in finding one parameter $s \in \mathcal{S}$ that have the best theoretical cost $\mathbb{E}[\gamma_n(s)]$. The purpose of M-estimation is exactly to define a random point that we hope to be close.

Definition 8. *We define the following notions.*

1. *Let s be the **target** parameter defined as*
$$s \in \operatorname{argmin}_{t\in\mathcal{S}} \mathbb{E}[\gamma_n(t)].$$
2. *We define the **M-estimator** based on the risk function as*
$$\hat{s} \in \operatorname{argmin}_{t\in\mathcal{S}} \gamma_n(t).$$
3. *The **cost** of choosing the parameter t is given by*
$$R(t) = \mathbb{E}[\gamma_n(t)]$$
*and the **risk** of the estimator is the quantity $R(\hat{s})$.*

It has to be stated somewhere that a M-estimator $\hat{s}$ is, obviously, depending on the *set of parameters* $\mathcal{S}$ and of the *form of the random criteria* γ_n

Empirical Measure Most of the time, the criteria $\gamma_n(t)$ can be rewritten in the setting of empirical processes where a sum of independent terms is considered. For any measure μ and any function $f : \mathcal{X} \mapsto \mathbb{R}$ integrable with respect to μ, we define

$$\mu f = \mu(f) = \int_{\mathcal{X}} f d\mu.$$

Obviously, for any function $f : \mathcal{X} \mapsto \mathbb{R}$ integrable with respect to P, we have

$$\begin{aligned}Pf &= \mathbb{E}[f(X)] \\ P_n f &= \frac{1}{n}\sum_{i=1}^{n} f(X_i)\end{aligned}$$

where $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is called the **empirical measure**.

10.2 Examples

In the following examples, we remark that the random criteria $\gamma_n(t)$ takes the specific form $P_n f_t$ for a good choice of f_t.

Empirical mean and empirical median Estimating the mean and the median of a sample of random vectors taking values in a set $\mathcal{X} = \mathbb{R}^k$ can be seen as a problem of M-estimation. The parameters are also elements of $\mathbb{R}^k$ then we set $\mathcal{S} = \mathbb{R}^k$.

- When $\gamma_n(t) = P_n f_t$ where $f_t(x) = (y-t)^2$, the target parameter s is simply the expected value $\mathbb{E}[X]$.
- When ones uses $f_t(x) = |y-t|$, the minimizer is just the (a) median of X.

Exercice 18. *Show that the minimum of* $\mathbb{E}\left[(Y-t)^2\right]$ *is attained for* $t = \mathbb{E}[Y]$ *and show that* $\mathbb{E}[|Y-t|]$ *is attained for* $t = Med(Y)$.

Least square regression In this context, we assume that the space $\mathcal{X}$ takes the form $\mathcal{X} = \mathcal{Z} \times \mathbb{R}$ for a measurable space $\mathcal{Z}$ and that $X = (Z, Y) \in \mathcal{Z} \times \mathbb{R}$ of law P and such that

$$Y = m(Z) + \sigma(Z)\varepsilon,$$

with $\mathbb{E}\left[Y^2\right] < \infty$ and $\sigma(Z) \geq 0$. The noise term ε is suppose to be independent of Z and standardized (i.e. $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}\left[\varepsilon^2\right] = 1$).

- The set of parameters is $\mathcal{S} = L_2(P) := \{s : \mathcal{Z} \to \mathbb{R} \ ; \ \mathbb{E}\left[s^2(Z)\right] < \infty\}$.
- The cost function is $\gamma_n(t) = P_n f_t$ where $f_t(x) = (y - t(z))^2$.
- The target $m : z \mapsto \mathbb{E}[Y|Z=z]$ is called the **regression function** of Y by Z.

The estimator $\hat{s}$ is the least square estimator (LSE).

Binary classification The binary classification deals with the problem of labeling a random variable Z by a number 0 or 1. The data points are, then, of the form $X_i = (Z_i, Y_i)$ where $Z_i \in \mathcal{Z}$ and $Y_i \in \{0,1\}$. Then $\mathcal{X} = \mathcal{Z} \times \{0,1\}$ and,

- The set of parameters is $\mathcal{S} = \{s : \mathcal{X} \to \{0,1\} \text{ measurable}\}$.
- The cost function is $\gamma_n(t) = P_n f_t$ where $f_t(x) = \mathbb{1}_{y \neq t(z)}$.
- The target $s_*(z) = \mathbb{1}_{\mathbb{E}[Y|Z=z] \geq 1/2}$ is called **Bayes classifier**.

The estimator $\hat{s}$ is the binary classifier.

Maximum likelihood We assume that X has a density f with respect to a measure μ,

$$f = \frac{dP}{d\mu}$$

and that $(\log f)_+$ is integrable with respect to P. Then:

- The set of parameters is $\mathcal{S} = \{s : \mathcal{X} \to \mathbb{R}_+ \ ; \ \int_{\mathcal{X}} s d\mu = 1 \text{ and } P(\log s)_+ < \infty\}$.
- The cost function is $\gamma_n(t) = P_n f_t$ where $f_t(x) = -\log(s(x))$.
- The target f is the density of X.

The estimator $\hat{s}$ is then the maximum likelihood estimator (MLE).

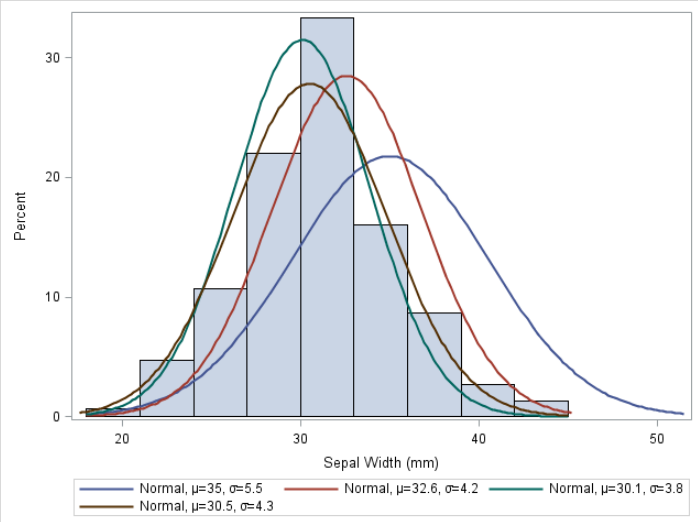


Figure 10.1: An example of MLE done by hands.

10.3 Theoretical study

For simplicity, we derive the following study in the context seen above, where the cost function $\gamma_n(t)$ takes the form of $P_n f_t$. The choice of the form of the function f_t depends on the statistical context. Hence the theoretical cost $\mathbb{E}[\gamma_n(t)]$ takes the form Pf_t. When one wants to study the deviation between the target s defined as the minimizer of Pf_t and the M-estimator $\hat{s}$ defined as the minimizer of $P_n f_t$, it is a good idea to control the difference $P_n f_t - Pf_t$. This enters naturally in the context of empirical process theory.

Definition 9. *Let $\mathcal{F}$ be a subset of $L_1(P)$. The functional*

$$\begin{aligned}\Phi\colon \mathcal{F} &\to \mathbb{R}\\ f &\mapsto P_n f - Pf\end{aligned}$$

also denoted $((P_n - P)f)_{f\in\mathcal{F}}$ is called the ***empirical process*** *over the class $\mathcal{F}$.*

This point of view is the one taken by numerous authors for a general study of M-estimators on metric sets of parameters. The interested reader is advised to take a look at [13], [14] or [8].

10.3.1 Consistency of M-estimators

Bounding the excess risk As defined earlier, the quality of the M-estimator is measured by its risk $R(\hat{s})$. A first step to prove the consistency of the estimator $\hat{s}$ is to control the so-called **excess risk**

$$R(\hat{s}) - R(s).$$

The convergence towards 0 of $R(\hat{s}) - R(s)$ is not directly linked to the convergence of $\hat{s}$ towards s. Indeed, if the function R has numerous local minimum then tracking the convergence of $\hat{s}$ becomes hard even though one has $R(\hat{s}) - R(s) \to 0$ as $n \to +\infty$. In the literature, many author do not bridge this step and only look for the asymptotic behavior of the excess risk of the estimator. If one wants to overcome this issue, several leads are possible. The most common one may be to assume convexity or strong convexity.

Definition 10 (Strong convexity)**.** *Let $\mu > 0$, U be a convex open subset of $\mathbb{R}^k$ and $f : U \subset \mathbb{R}^k \to \mathbb{R}$ be a differentiable function. We say that a function is μ-strongly convex if one of the following equivalent conditions is verified.*

1. $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|x-y\|^2$ *for any* $x, y \in U$.
2. *The function* $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ *is convex.*
3. $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \mu\|x-y\|$.

Exercice 19. *Prove the equivalences in Definition 10.*

The equivalences in Definition 10 still hold when f is assumed to have sub-gradients. See the details in [3, Section 9.1.2]. The other option is to assume that for a distance d defined on the set $\mathcal{S}$ of parameters, we have

$$\eta d(t,s)^2 \leq R(t) - R(s), \quad \forall t \in \mathcal{S}$$

for a positive constant η. The power 2 in the previous inequality is arbitrary but is often chosen in the literature. In the sequel, we do not comment more on this fact and focus on proving consistency results only for the excess risk $R(\hat{s}) - R(s)$. The following lemma encodes a crucial decomposition of the risk.

Lemma 18. *Let $\forall t \in \mathcal{S}$, $R_n(t) = \gamma_n(t)$ and assume that is satisfies*

$$\sup_{t \in \mathcal{S}} |R_n(t) - R(t)| \xrightarrow{\mathbb{P}} 0,$$

then $R(\hat{s}) - R(s) \xrightarrow{\mathbb{P}} 0$.

Proof. We have

$$\begin{aligned}
0 \leq R(\hat{s}) - R(s) & \\
&= [R(\hat{s}) - R_n(\hat{s})] + [R_n(\hat{s}) - R_n(s)] + [R_n(s) - R(s)] \\
&\leq [R(\hat{s}) - R_n(\hat{s})] + [R_n(s) - R(s)] \\
&\leq 2 \sup_{t \in \mathcal{S}} |R_n(t) - R(t)| \xrightarrow{\mathbb{P}} 0
\end{aligned}$$

□

Chapter 11

Model Selection

We are in the context when the quantity to estimate is some complex object such a graph, a function, etc... If we take the case of density estimation as a generic example for the context, one has to determine a objective function inside a possibly enormous set of functions (think to all continuous function from $\mathbb{R}$ to $[0,1]$ for example). Hence, a natural strategy is to reduce the set of possible solution at the price of possibly deteriorating the quality of the estimation. We put it in context in the following. This chapter is inspired from the thesis of Adrien Saumard [10].

11.1 Introduction

Let $X_1, \dots, X_n$ be i.i.d random variables taking values in a set $\mathcal{X}$. Let $\mathcal{S}$ be a set (possibly very complex) of parameters (DEFINE this). We also define a random criteria γ_n sometimes referred as contrast as a function of the data for measuring the quality (DEFINE that) of a parameter $t \in \mathcal{S}$. More concretely, let

$$\begin{aligned} \gamma_n \colon \mathcal{S} &\to \mathbb{R}_+^* \\ t &\mapsto \gamma_n(t) := \gamma_n(X_1, \dots, X_n, t) \end{aligned}$$

be the *cost* (or *risk*) function. In many cases, the cost function takes the form of a sum of independent random quantities $\gamma_n(t) = n^{-1} \sum_i c(X_i, t)$ in such a way that $\gamma_n(t)$ can be rewritten in the context of empirical processes theory $\gamma_n(t)$ (see Definition REF). We, now, introduce the important vocabulary in the setting of model selection.

Definition 11. *We define the following notions.*

1. *The* ***empirical cost*** *for a parameter $t \in \mathcal{S}$ is $\gamma_n(t)$.*
2. *The* ***cost*** *or* ***risk*** *is $\mathbb{E}[\gamma_n(t)]$.*
3. *A subset $S \subset \mathcal{S}$ is called a* ***model****. When one has access to a class of such subsets $(S_m)_{m \in \mathcal{M}}$, we also call model the index m of the model S_m.*
4. *Let s be the* ***target*** *parameter defined as*
$$s \in \operatorname{argmin}_{t \in \mathcal{S}} \mathbb{E}[\gamma_n(t)] .$$
It is the theoretical benchmark for the problem of optimizing the cost. For each model $m \in \mathcal{M}$ we define the ***projected target*** *as a minimizer of the cost on the model S_m,*
$$s_m \in \operatorname{argmin}_{t \in S_m} \mathbb{E}[\gamma_n(t)] .$$
5. *For each model m, we define the associated* ***M-estimator*** *based on the risk function as*
$$\hat{s}_m \in \operatorname{argmin}_{t \in S_m} \gamma_n(t).$$
6. *Finally, among the models $\mathcal{M}$ we choose the* ***optimal*** *model for which the cost of its M-estimator is minimal,*
$$m_* \in \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\gamma_n(\hat{s}_m)] \tag{11.1}$$

Question: If one have access to a class of models $(S_m)_{m \in \mathcal{M}}$, how can one choose a model m and an estimator $\tilde{s}$ as an element of the model S_m such that it is a good estimator of s?

Selecting among M-estimators In the definitions *5.* and *6.*, we reduced the diversity of estimator we consider. Indeed, we only assume that we construct a M-estimator corresponding to each model S_m. As a result, the estimator $\tilde{s}$ is to be chosen among the family $(\hat{s}_m)_{m\in\mathcal{M}}$ as we will develop further in the following.

Target model The model m_* or S_{m_*} give an associated estimator $\hat{s}_{m_*}$ having the best theoretical performance (in the sense of (11.1)) among the class of M-estimators $(\hat{s}_m)_{m\in\mathcal{M}}$. In that sense, s_{m_*} is the best estimator to estimate the target s. However, it is not rigorously an estimator since it still depends on some parameter of the problem through m_*. This comes from that the minimization in (11.1) uses the true mean operator.

Avoiding a confusion : $\mathbb{E}_\gamma[\cdot]$ versus $\mathbb{E}[\cdot]$ In the following, we will have to distinguish between two kind of alea. The empirical cost is one source of randomness and an estimator in some model $\mathcal{M}$ gives another source. We attract the attention of the reader on the fact that as a function of a (non-random) parameter t, $\mathbb{E}[\gamma_n(t)] =: \mathbb{E}_\gamma[\gamma_n(t)]$ is no more random. Then, when one considers a estimator $\hat{s}_m$,

$$\begin{aligned} \mathbb{E}_\gamma[\gamma_n(\hat{s}_m)] &\qquad \text{is a random variable} \\ \mathbb{E}[\gamma_n(\hat{s}_m)] &\qquad \text{is a deterministic number} \end{aligned}$$

since the second quantity is simply the expected value of the random variable $\gamma_n(\hat{s}_m)$. The reader has to be careful that we do **not** have $\mathbb{E}[\mathbb{E}_\gamma[\gamma_n(\hat{s}_m)]] = \mathbb{E}[\gamma_n(\hat{s}_m)]$ but we obviously have that $\mathbb{E}_\gamma[\gamma_n(t)] = \mathbb{E}[\gamma_n(t)]$ for any deterministic point $t \in \mathcal{S}$.

Loss functions In order to quantify the goodness of an estimator, one has to define a non-negative quantity that quantifies the gap between an estimated parameter and s. In the literature, there are two natural and common choices. We define the **deterministic loss** function ℓ_{det} of an estimator $\tilde{s}$ around the target point s by

$$\ell_{det}(\tilde{s}, s) = \mathbb{E}[\gamma_n(\tilde{s})] - \mathbb{E}[\gamma_n(s)].$$

We define the **random loss** function ℓ_{ran} as

$$\ell_{ran}(\tilde{s}, s) = \mathbb{E}_\gamma[\gamma_n(\tilde{s})] - \mathbb{E}[\gamma_n(s)].$$

In each section, we specify which loss is considered and we will use the generic notation ℓ for both cases since there will not be confusion. Note that, for both cases, the projected target s_m is a minimizer on S_m of $\ell(t, s)$. At this point, it is clear that a model S_m too "small" is not likely to embed properly the problem as the target s will be far from its closest point in S_m and then one has to look for a rich enough model to hope to get a good estimator $\tilde{s}$ of the target.

Over-fitting At first sight, the question seems to be answered by a direct minimization of the empirical cost by

$$\hat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \gamma_n(\hat{s}_m) \tag{11.2}$$

which will have the tendency to always choose the biggest (in the sense of inclusion) model S_m among the possibilities. However, a "big" model have the tendency to suffer a negative bias. Indeed, calling $\overline{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}_\gamma[\gamma_n(t)]$ and using

$$\mathbb{E}_\gamma[\gamma_n(\hat{s}_m)] = \mathbb{E}_\gamma[\gamma_n(s_m)] + \underbrace{\mathbb{E}_\gamma[\gamma_n(\hat{s}_m) - \gamma_n(s_m)]}_{\geq 0}$$

where the operator $\mathbb{E}_\gamma$ only operates on γ_n and not on $\hat{s}_m$, and

$$\gamma_n(\hat{s}_m) = \gamma_n(s_m) - \underbrace{(\gamma_n(s_m) - \gamma_n(\hat{s}_m))}_{\geq 0}$$

one can write $\overline{\gamma}_n(\hat{s}_m) = \overline{\gamma}_n(s_m) - (\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\hat{s}_m))$. Since the point s_m is not random, $\overline{\gamma}_n(s_m)$ is centered (or without bias). Nevertheless, the term $\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\hat{s}_m)$ is non-negative and then

$$\mathbb{E}[\overline{\gamma}_n(\hat{s}_m)] \leq 0. \tag{11.3}$$

This can be interpreted as the fact that the minimization in (11.2) introduces a negative bias so that $\gamma_n(\hat{s}_m)$ is too small compared to its cost $\mathbb{E}[\gamma_n(\hat{s}_m)]$. This occurs in the over-fitting phenomena using a model with too much details/parameters.

Practice 1. *BUILD AN EXAMPLE TO COMPUTE OVERFITTING*

Hence the term that control the bias of the over-fitted estimator is $\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\hat{s}_m)$. This bias is controlled by the complexity (the richer the more complex), of the model m chosen to build the estimator.

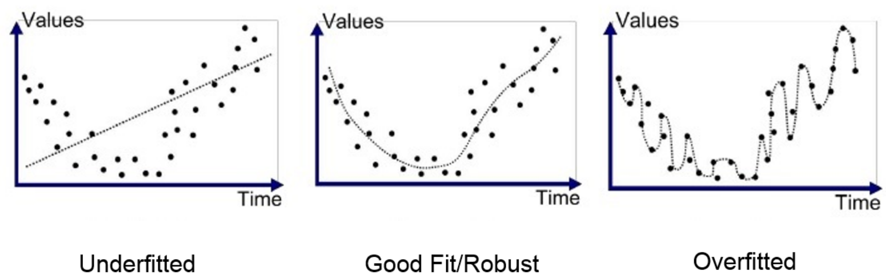


Figure 11.1: A typical problem of underspecified (left) vs adapted (center) vs over-fitting (right)

11.1.1 A solution through penalization

A solution to overcome the issue of over-fitting (negative bias) is to correct the estimator by a slightly modified minimization by adding a term of penalization of a model.

Definition 12. *A penalization on the class of models* $(S_m)_{m\in\mathcal{M}}$ *is a function* $\text{pen} : \mathcal{M} \to \mathbb{R}_+$*. We allow* $\text{pen}(m)$ *to be a random variable depending on the data* $X_1, \ldots, X_n$*.*

The new estimator is then defined as a minimPseizer of

$$\hat{m} \in \text{argmin}_{m\in\mathcal{M}}\{\gamma_n(\hat{s}_m) + \text{pen}(m)\}. \tag{11.4}$$

For clarity in the notations, from now on, we denote by $\tilde{s}$ the **selected** estimator using (11.4) estimator $\hat{s}_{\hat{m}}$.

Ideal penalizations

We define the ideal penalizations

$$\text{pen}^{id}_{det}(m) = \mathbb{E}\left[\gamma_n(\hat{s}_m)\right] - \gamma_n(\hat{s}_m) \tag{11.5}$$
$$\text{pen}^{\text{id}}_{ran}(m) = \mathbb{E}_\gamma\left[\gamma_n(\hat{s}_m)\right] - \gamma_n(\hat{s}_m) = -\overline{\gamma}_n(\hat{s}_m). \tag{11.6}$$

In practice, pen^{id} cannot be used to tune the estimator since it depends on theoretical quantities such that the true mean of $\gamma_n(\hat{s}_m)$. Assume for a second that we choose $\text{pen} = \text{pen}^{id}_{det}$, then it is clear that $\hat{m} = m_*$ and this choice would achieve the prefect estimator $\hat{s}_{m_*}$.

11.1.2 A good class of results: Oracle bounds

The purpose of this section is to define properly the form of the results that one may want to develop. One is usually interested in proving that the estimator in question satisfy the same kind of guaranties than the best estimator provided in the class $(S_m)_{m\in\mathcal{M}}$. We will give at least two different mathematical meaning of this sentence. Since the calculations on ℓ_{ran} and ℓ_{det} are similar, we will give a unified notation ℓ for both loss functions and denote by $\mathbf{E}[\cdot]$ the associated expectation that is either $\mathbb{E}$ or $\mathbb{E}_\gamma$ depending on the case.

Oracle bounds We will be looking for bounds of the form

$$\ell(\tilde{s}, s) \leq C \inf_{m\in\mathcal{M}} \ell(\hat{s}_m, s) + \text{Dev} = C\ell(\hat{s}_{m_*}, s) + \text{Dev} \tag{11.7}$$

for C a positive constant. A result as (11.7) is called an **oracle bound**. In other words, we ask that the desired estimator $\tilde{s}$ is not worse than a constant times the best theoretical choice $\hat{s}_{m_*}$. The Equation (11.1) has to be understood as a deterministic bound for ℓ_{det} and the term Dev is a deterministic deviation whereas, in the case ℓ_{ran}, the bound holds in *expectation* or *high probability* and the deviation term is allowed to be a random quantity. Oracle bounds can also take the form of

$$\ell(\tilde{s}, s) \leq C \inf_{m\in\mathcal{M}} (\ell(s_m, s) + \text{pen}(m)) + \text{Dev}' \tag{11.8}$$

where the infimum describes the best possible projection on S_m weighed by the penalization term.

A generic calculation We have, from (11.6), the following calculations

$$\begin{aligned}\ell(\tilde{s}, s) &= \mathbf{E}\left[\gamma_n(\tilde{s})\right] - \mathbf{E}\left[\gamma_n(s)\right] \\ &= \gamma_n(\tilde{s}) + \mathrm{pen}^{\mathrm{id}}(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \\ &= \gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \\ &\leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \end{aligned} \tag{11.9}$$

The next step concerns the bound on $\gamma_n(\hat{s}_m)$. It is actually possible to derive two kind of results that we detail in the next two paragraphs. Each strategy lead to different form of oracle bounds.

First solution The first solution is to write $\gamma_n(\hat{s}_m)$ as

$$\gamma_n(\hat{s}_m) = -\,\mathrm{pen}^{id}(m) + \mathbf{E}\left[\gamma_n(\hat{s}_m)\right].$$

and then

$$\ell(\tilde{s}, s) \leq \ell(\hat{s}_m, s) + (\mathrm{pen} - \mathrm{pen}^{id})(m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) \tag{11.10}$$

The goal of the penalization step is, then, to look for good approximation of the ideal penalization $\mathrm{pen}^{\mathrm{id}}$ by pen over the models $m \in \mathcal{M}$.

Second solution The second solution consists in bounding $\gamma_n(\hat{s}_m)$ in a direct manner thanks to the definition of the estimator $\hat{s}_m$. Starting again from (11.9) and using

$$\gamma_n(\hat{s}_m) \leq \gamma_n(s_m),$$

the bound on $\ell(\tilde{s}, s)$ becomes

$$\ell(\tilde{s}, s) \leq \ell(s_m, s) + \mathrm{pen}(m) + \overline{\gamma}_n(s_m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) \tag{11.11}$$

We see that when one is able to find a penalization close to the ideal penalization, one can hope to get an oracle inequality as (11.7). For example, if one can control uniformly the deviation between $\mathrm{pen}_{\mathrm{id}}$ and pen with high probability,

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \mathrm{pen}(m) \leq \mathrm{pen}_{\mathrm{id}}(m) + C \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s) \tag{11.12}$$

we get

$$\ell(\tilde{s}, s) \leq (1 + C) \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s)$$

with high probability. An ideal context is when one is able to define a penalization such that, with high probability,

$$|\,\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)| \leq \varepsilon \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s) \tag{11.13}$$

so that

$$\ell(\tilde{s}, s) \leq \frac{1+\varepsilon}{1-\varepsilon} \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s)$$

which is asymptotically optimal if $\varepsilon \to 0$ as $n \to \infty$.

Part III

Annexes

Chapter 12

Extra definitions

12.0.1 Sumable familly

Definition 13. *Let $(E, \|\cdot\|)$ a normed vector space. We say that a family $(a_i)_{i\in I}$ of elements of E is **sumable** if there exists an element S of E such that $\forall\varepsilon > 0$, $\exists J_\varepsilon$ a finite subset of I such that $\forall J$ finite $\subset I$,*

$$J \supseteq J_\varepsilon \implies \left\| \sum_{i\in J} a_i - S \right\| \leq \varepsilon.$$

Then S is unique and we call it the sum of the sumable family a_i.

Proposition 23. *If the elements a_i are non-negative, then*

$$(a_i)_{i\in I} \text{ is sumable } \Leftrightarrow \begin{array}{l} J_{>0} := \{i \in I : a_i \neq 0\} \text{ is at most countable} \\ \text{and the serie } \sum_{i\in J_{>0}} a_i \text{ is convergent.} \end{array}$$

Proof. Simply note that for any $\varepsilon > 0$, the set $\{i : a_i > 2\varepsilon\}$ is finite since it is included in J_ε. Then we have that

$$\{i : a_i \neq 0\} = \bigcup_{n\in\mathbb{N}} \{i : a_i > \frac{1}{n}\}.$$

□

This is actually possible to adapt the proof to get the result for the general sequence (a_i) where the result on the serie is that it is commutatively convergent i.e. that any permutation of the terms lead to the same sum.

12.0.2 Processes

Definition 14. *A **modification** of a process $(X_t)_{t\in\mathcal{T}}$ is a process $(\tilde{X}_t)_{t\in\mathcal{T}}$ such that*

$$\mathbb{P}\left(\forall t,\ X_t = \tilde{X}_t\right) = 1.$$

Chapter 13

Functional Analysis

13.1 Lemmas

We give here the proofs of some technical results.

Lemma 19. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be complex numbers such that $\forall i$, $|a_i| \leq 1$ and $|b_i| \leq 1$. Then*

$$|a_1 a_2 \dots a_n - b_1 b_2 \dots b_n| \leq \sum_{i=1}^{n} |a_i - b_i|.$$

Proof. It is possible to rewrite $a_1 a_2 \dots a_n - b_1 b_2 \dots b_n$ as

$$\begin{aligned} a_1 a_2 \dots a_n - b_1 b_2 \dots b_n = & a_1 a_2 \dots a_n - a_1 a_2 \dots a_{n-1} b_n \\ & + a_1 a_2 \dots a_{n-1} b_n - a_1 a_2 \dots a_{n-2} b_{n-1} b_n \\ & + \dots \\ & + a_1 b_2 \dots b_n - b_1 b_2 \dots b_n \end{aligned}$$

Then

$$|a_1 a_2 \dots a_n - b_1 b_2 \dots b_n| \leq |a_n - b_n| + \dots + |a_1 - b_1|$$

since the complex numbers are all of modulus less or equal to 1. □

Lemma 20. *For any pair of positive numbers a and b, we have that for any $p \geq 1$,*

$$(a+b)^p \leq 2^{p-1}(a^p + b^p)$$

Proof. Use the convexity of $x \mapsto x^p$ between the points a and b with $\lambda = 1 - \lambda = 1/2$. □

Lemma 21. *For any complex z such that $\Re(z) \leq 0$, we have*

$$|e^z - 1 - z| \leq \frac{|z|^2}{2}$$

Proof. By the Taylor-Young formula, we see that

$$|e^z - 1 - z| = \left| \int_0^1 (t-1) z^2 e^{tz} dt \right| \leq |z|^2 \int_0^1 (1-t) dt = \frac{|z|^2}{2}$$

where we used that $|e^{tz}| \leq 1$, by the fact that $\Re(z) \leq 0$. □

Lemma 22. *Let I be an open interval of $\mathbb{R}$ and let $c : I \to \mathbb{R}$ be a convex function. Then we have the following facts*

a) c is continuous on I.

b) For all $x \in I$, c has a left derivative $c'_l(x)$ and a right derivative $c'_r(x)$ such that $c'_l(x) \leq c'_r(x)$.

c) Fix any $v \in I$ then for all $D \in [c'_l(v), c'_r(v)]$, we have that $\forall x \in I$, $c(x) \geq D(x-v) + c(v)$.

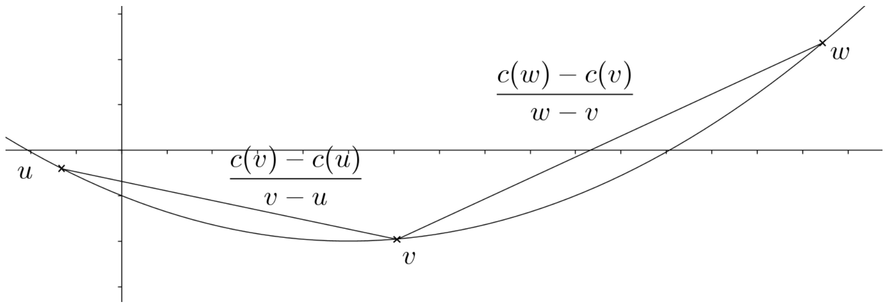


Figure 13.1: Inequality (13.1)

d) There exists two sequences $(a_n)_n$ and $(b_n)_n$ of reals such that

$$\forall x \in I, \quad c(x) = \sup_n (a_n x + b_n).$$

Proof. If one takes $u < v < w$ elements of I, we have that

$$\Delta_{u,v} \leq \Delta_{u,w} \leq \Delta_{v,w} \quad \text{where} \quad \Delta_{x,y} = \frac{c(y) - c(x)}{y - x}. \tag{13.1}$$

It is obvious to deduce that $\Delta_{x,y}$ is increasing in both x and y. Now let $v_0 \in (u, w)$ fixed. So

$$|C(v) - C(v_0)| = |\Delta_{v_0,v}||v - v_0| \leq \max\{|\Delta_{v_0,w}|, |\Delta_{u,v_0}|\}|v - v_0| \underset{v \to v_0}{\longrightarrow} 0$$

and $a)$ is proved. From (13.1), we prove that

$$c_l'(v) = \lim_{u \uparrow v} \Delta_{u,v} \leq \lim_{w \downarrow v} \Delta_{v,w} = c_r'(v).$$

The limits exists since the limits are defined for increasing (resp. decreasing) and upper bounded (resp. lower bounded) functions. Let $D \in [c_l'(v), c_r'(v)]$ and let $x \in I$. If $x \geq v$, we have that $D \leq c_r'(v) \leq \Delta_{v,x} = (c(x) - c(v))/(x - v)$. The case $x\ lev$ is obtained symmetrically. To prove $d)$, we consider the point $c)$ for all $q \in I \cap \mathbb{Q}$ where we choose for example $D_q = (c_l'(q) + c_r'(q))/2$ and we define

$$f(x) = \sup_{q \in I \cap \mathbb{Q}} (D_q(x - q) + c(q)).$$

Now by density one can choose $(q_n)_n$ a sequence of rationals in I such that $q_n \to x$. Then,

$$c(x) = \lim_{n \to \infty} (D_{q_n}(x - q_n) + c(q_n)) \leq \sup_{q \in I \cap \mathbb{Q}} (D_q(x - q) + c(q)) = f(x) \leq c(x).$$

We have $c = f$ and since $I \cap \mathbb{Q}$ is countable, one can renumerate the elements in a sequence. □

13.2 Basic facts on integrable functions

Proposition 24. *Let $f \geq 0$ be an integrable function, then for any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\forall F \in \mathcal{B}(\mathbb{R}),\ \mathbb{P}(F) \leq \delta \implies \int f(x) \mathbb{1}_{f(x) \in F} \leq \varepsilon.$$

Proof. Assume that the conclusion is false, then, there exist ε_0 and a sequence of sets $(F_n)_n$ such that

$$\mathbb{P}(F_n) \leq 2^{-n} \quad \text{and} \quad \int f(x) \mathbb{1}_{f(x) \in F_n} > \varepsilon_0.$$

Defining, $F = \limsup F_n$, we get from Borel-Cantelli lemma that $\mathbb{P}(F) = 0$. However, reverse Fatou lemma shows that

$$\int f(x) \mathbb{1}_{f(x) \in F} > \varepsilon_0$$

but this is impossible since the integration of over a event of probability 0 is always 0. The absurdity of the assumption gives the result. □

Corollary 6. *Let $f \geq 0$ be an integrable function, then*

$$\int f(x) \mathbb{1}_{|f(x)| > t} dx \underset{t \to \infty}{\longrightarrow} 0.$$

13.3 Basic properties and Fourier transform

Fact 1. *The convolution between two measures given by*

$$\mu \star \nu(A) = \int_{\mathbb{R}^k \times \mathbb{R}^k} \mathbb{1}_{x+y \in A} d\mu(x) d\nu(y)$$

is a probability measure.

Proof. Obviously, $\mu \star \nu(\mathbb{R}^k) = 1$. Let $A_1, \dots, A_n, \dots$ be a countable family of disjoints elements of the borelian σ-algebra. Then one has that

$$\mathbb{1}_{\cup_{i \geq 1} A_i} = \sum_{i \geq 1} \mathbb{1}_{A_i}$$

which implies $\mu \star \nu(\cup_{i\geq 1} A_i) = \sum_{i \geq 1} \mu \star \nu(A_i)$, by linearity of the integral. □

We recall proposition 7.

Proposition 25. *For μ and ν two probability measures,*

- $\|\mathcal{F}\mu\|_\infty \leq 1$.
- $\mathcal{F}(\mu \star \nu) = (\mathcal{F}\mu) \times (\mathcal{F}\nu)$.

Proof. The first fact is obvious since the integrand has a modulus bounded by 1. For the second point, we see that for any integrable function f,

$$\int_{\mathbb{R}^k} f(z) d(\mu \star \nu)(z) = \iint_{\mathbb{R}^k \times \mathbb{R}^k} f(x+y) d\mu(x) d\nu(y).$$

This can be seen by approximation of positive functions by simple functions. Then

$$\begin{aligned}
\mathcal{F}(\mu \star \nu)(\xi) &= \int_{\mathbb{R}^k} \exp(-iz \cdot \xi) d(\mu \star \nu)(z) \\
&= \iint_{\mathbb{R}^k \times \mathbb{R}^k} \exp(-i(x+y) \cdot \xi) d\mu(x) d\nu(y) \\
&= \left(\int_{\mathbb{R}^k} \exp(-iz \cdot \xi) d\mu(z) \right) \left(\int_{\mathbb{R}^k} \exp(-iy \cdot \xi) d\nu(y) \right) \\
&= (\mathcal{F}\mu(\xi))(\mathcal{F}\nu(\xi))
\end{aligned}$$

□

Modulus of continuity Let $g : \mathbb{R}^k \to \mathbb{R}$ be a function. Its **modulus of continuity** $w(g, x, \delta)$ in x is a function taking values in $[0, +\infty]$ defined by

$$w(g, x, \delta) = \sup_{y \in \mathbb{R}^k : \|x - y\| \leq \delta} |g(y) - g(x)|.$$

By definition, it can be seen that

$$g \text{ is continuous at } x \Leftrightarrow \lim_{\delta \to 0} w(g, x, \delta) = 0.$$

Regularizing sequence We say that a sequence $(\phi_n)_{n \in \mathbb{N}}$ of functions on $\mathbb{R}^k$ is a **regularizing sequence** if

1. For all n, $\phi_n \geq 0$.
2. For all n, $\int_{\mathbb{R}^k} \phi_n(x) dx = 1$.
3. For every $\varepsilon > 0$, $\int_{B(0,\varepsilon)^c} \phi_n(x) dx \underset{n \to \infty}{\longrightarrow} 0$.

Proposition 26. *Let $1 \leq p, q < \infty$ such that $p^{-1} + q^{-1} = 1$. Let ϕ_n be a regularizing sequence of functions in $L_q(\mathbb{R}^k)$. Then, for any $f \in L_p(\mathbb{R}^k)$, we have that*

$$f \star \phi_n \underset{n \to \infty}{\longrightarrow} f \quad (\text{in } L_p(\mathbb{R}^k)).$$

To prove that fact, we begin with a stronger case.

Lemma 23. *For a function g in $L_\infty(\mathbb{R}^k)$ continuous at x, we get*

$$g \star \phi_n(x) \underset{n \to \infty}{\longrightarrow} g(x)$$

Proof. Using the fact that ϕ_n is of total mass 1 by definition, we can write for any $\delta > 0$,

$$\begin{aligned} g \star \phi_n(x) - g(x) &= \int_{\mathbb{R}^k} [g(x-y) - g(x)]\phi_n(y)dy \\ &= \int_{B(0,\delta)} [g(x-y) - g(x)]\phi_n(y)dy + \int_{B(0,\delta)^c} [g(x-y) - g(x)]\phi_n(y)dy \\ &\leq w(g,x,\delta) + 2\|g\|_\infty \int_{B(0,\delta)^c} \phi_n(y)dy \end{aligned}$$

Now, by continuity, take $\delta > 0$ sufficiently small to get $w(g,x,\delta) \leq \varepsilon/2$ and then take n large enough to have the second term smaller than $\varepsilon/2$ as well. This finishes the proof. □

We are now able to prove Proposition 26.

Proof of Proposition 26. Since the family of regularizing functions ϕ_n are in $L_q(\mathbb{R}^k)$, the functions $f \star \phi_n$ are well defined. Then by Jensen's inequality,

$$|(f \star \phi_n)(x) - f(x)| \leq \int_{\mathbb{R}^k} |f(x-y) - f(x)|^p \phi_n(y)dy.$$

Integrating in x both sides and using Fubini's theorem (everything is positive) we get that

$$\|(f \star \phi_n) - f\|_p^p \leq \int_{\mathbb{R}^k} \|f_y - f\|_p^p \; \phi_n(y)dy, \tag{13.2}$$

where f_y holds for the function $x \mapsto f(x-y)$. Define $g(y) = \|f_y - f\|_p^p$, then it is a continuous bounded function such that $g(0) = 0$. Hence, looking at the right and side of Equation (13.2) as $g \star \phi_n(0)$ we get, by Lemma 23, that it converges to 0 as $n \to +\infty$. □

13.4 Distribution functions and simple functions

Definition 15. *A simple function is a function f such that there exists a finite number n of real values $\lambda_1, \ldots, \lambda_n$ and of measurable sets $A_1, \ldots, A_n$ such that*

$$f = \sum_{i=1}^n \lambda_i \mathbb{1}_{A_i}$$

Definition 16. *A function defined on an finite interval $I = [a,b]$ is said to be absolutely continuous on I, if $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $\forall n$ and every finite familly of intervals $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots, (\alpha_n, \beta_n)$ in I such that*

$$\sum_{i=1}^n (\beta_i - \alpha_i) < \delta,$$

we have,

$$\sum_{i=1}^n |f(\beta_i) - f(\alpha_i)| < \varepsilon$$

This definition implies the important,

Theorem 13. *Let $I = [a,b]$ and $f : I \to \mathbb{R}$ a non decreasing and absolutely continuous function. Then, f is almost surely differentiable on I, is in $L_1(\mathbb{R})$ and*

$$f(x) - f(a) = \int_a^x f'(t)dt \qquad \forall x \in [a,b].$$

Proof. This can be found in Rudin [9, Theorem 7.18] □

We have the useful lemma:

Lemma 24. *Let μ be a probability measure on $\mathcal{X}$ and let $f :\to [0,+\infty]$ a measurable function. Let $\phi : [0,+\infty) \to [0,+\infty]$ be a monotone function, absolutely continuous on $[0,T]$ for any $T < +\infty$ and such that $\phi(0) = 0$, then*

$$\int_{\mathcal{X}} (\phi \circ f)d\mu = \int_0^{+\infty} \mu\{f > t\}\phi'(t)dt \tag{13.3}$$

Proof. Since ϕ is absolutely continuous, it is almost surely differentiable. Now take a simple function f defined on $\mathcal{X}$ and let $E^t = \{x \in \mathcal{X} : f(x) > t\}$. The set E^t is measurable since it is a finite union of rectangles, then

$$\mu\{f > t\} = \mu(E^t) = \int_{\mathcal{X}} \mathbb{1}_{f(x)>t} d\mu(x)$$

and, by Fubini,

$$\int_0^{+\infty} \mu\{f > t\}\phi'(t)dt = \int_{\mathcal{X}} d\mu(x) \int_0^{+\infty} \mathbb{1}_{f(x)>t}\ \phi'(t)dt.$$

But the right hand side integral can be re-written in

$$\int_0^{+\infty} \mathbb{1}_{f(x)>t}\ \phi'(t)dt = \int_0^{f(x)} \phi'(t)dt = \phi(f(x)).$$

We end the proof by a classical density argument to insure the validity of (13.3) for any measurable function. □

A special case of Lemma 24 is the following result.

Corollary 7. *For any non negative random variable X,*

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(X > t)\, dt.$$

We draw the attention of the reader to the fact that the integral can also be written

$$\int_0^{+\infty} \mathbb{P}(X \geq t)\, dt \tag{13.4}$$

since integration on the open $(0, +\infty)$ or on $[0, +\infty)$ are equivalent for the Lebesgue measure dt.

Proof. Apply Lemma 24 for f, ϕ both equal to the identity function. □

13.5 Dominated convergence theorem

We recall rapidly the dominated convergence theorem that we reduce into (DOM) anywhere else in the notes. In the sequel of this section, we denote by $L_1(\mathcal{X}, \mu)$ the set of integrable functions on the measure space $(\mathcal{X}, \mu)$. When convenient, we adopt the notation

$$\mu(f) = \int_{\mathcal{X}} f(x) d\mu(x).$$

13.5.1 Dominated convergence

Theorem 14 (Dominated convergence (DOM))**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions. Assume that for any $x \in \mathcal{X}$, $f_n(x) \to f(x)$ for f a measurable function. Assume also that there exists a non-negative function $g \in L_1(\mathcal{X}, \mu)$ such that,*

$$|f_n(x)| \leq g(x), \quad \forall x \in \mathcal{X}, \forall n \in \mathbb{N}.$$

Then,

$$f_n \xrightarrow{\mathbb{L}_1} f \text{ in } L_1(\mathcal{X}, \mu),$$

and then

$$\int_{\mathcal{X}} f_n(x) d\mu(x) \underset{n\to\infty}{\longrightarrow} \int_{\mathcal{X}} f(x) d\mu(x)$$

Proof. This theorem is a direct consequence of Fatou's lemma. Taking the limit in inequations, we see that $|f| \leq g$ and then $|f_n - f| \leq 2g$. The reverse Fatou Lemma gives

$$\limsup \mu(|f_n - f|) \leq \mu(\limsup |f_n - f|) = \mu(0) = 0.$$

This implies the convergence L_1. Then, by Jensen inequality,

$$|\mu(f_n) - \mu(f)| \leq \mu(|f_n - f|) \underset{n\to\infty}{\longrightarrow} 0.$$

□

One can notice that the only tool used in the proof is the reverse Fatou lemma. It is then immediate to show the following corollary that study the case of convergence in probability of the sequence of integrated random variables.

Corollary 8 (Dominated convergence ($\mathbb{P}$ version)). *Let $(X_n)_n$ be a sequence of random functions such that for any x, $X_n(x) \xrightarrow{\mathbb{P}} X(x)$ and such that there exists a random function Y, integrable with respect to a measure μ such that $\forall n$, $|X_n| \leq Y$. Then*

$$\int X_n(x) d\mu(x) \xrightarrow{\mathbb{P}} \int X(x) d\mu(x).$$

Proof. We follow the proof of Theorem 14 with the additional use of reverse Fatou lemma,

$$\limsup \mathbb{P}\left(\mu(|X_n - X|) \geq \varepsilon\right) \leq \mathbb{P}\left(\limsup \mu(|X_n - X|) \geq \varepsilon\right) \leq \mathbb{P}\left(\mu(\limsup |X_n - X|) \geq \varepsilon\right) = \mathbb{P}\left(\mu(0) \geq .\varepsilon\right) = \mathbb{P}\left(0 \geq \varepsilon\right) = 0.$$

The reader may be confused by the first inequality. We used reverse Fatou for the functions $\mathbb{1}_{\mu(|X_n - X|) \geq \varepsilon}$ and the fact that for any sequence of random variables (Z_n),

$$\limsup \mathbb{1}_{Z_n \geq \varepsilon} = \lim_{n \to \infty} \sup_{k \geq n} \mathbb{1}_{Z_k \geq \varepsilon} = \lim_{n \to \infty} \mathbb{1}_{\sup_{k \geq n}(Z_k) \geq \varepsilon} = \mathbb{1}_{\lim_{n \to \infty} \sup_{k \geq n}(Z_k) \geq \varepsilon}$$

where the last equality is clear since the sequence $(\sup_{k \geq n}(Z_k))_n$ is monotone. □

Lemma 25 (Scheffé). *Assume that f_n and f are non-negative functions in $L_1(\mathcal{X}, \mu)$ and suppose that $f_n \to f$ a.e. Then*

$$\int |f_n - f| d\mu \underset{n \to \infty}{\longrightarrow} 0 \text{ if and only if } \int f_n d\mu \underset{n \to \infty}{\longrightarrow} \int f d\mu$$

Proof. The direct sense is obvious. For the reverse, assume that

$$\mu(f_n) \underset{n \to \infty}{\longrightarrow} \mu(f).$$

First, one can notice that $(f_n - f)^- \leq f - f_n \leq f$ by non-negativity of f_n and then (DOM) implies that $\mu((f_n - f)^-) \to 0$. For the positive part,

$$\mu((f_n - f)^+) = \mu((f_n - f)\mathbb{1}_{f_n \geq f}) = \mu(f_n) - \mu(f) - \mu((f_n - f)\mathbb{1}_{f_n < f})$$

and $|\mu((f_n - f)\mathbb{1}_{f_n < f})| \leq |\mu((f_n - f)^-)| \to 0$. Then, $\mu((f_n - f)^+) \to 0$ and

$$\mu(|f_n - f|) = \mu((f_n - f)^+) + \mu((f_n - f)^-) \to 0.$$

□

Scheffé Lemma have an important consequence for density functions associated with a probability measure P.

Corollary 9. *The almost sure convergence of densities imply convergence in $L_1(\mathcal{X}, P)$.*

Proof. Use Scheffé Lemma with the 'if' part since $\forall n,\ P(f_n) = 1 = P(f)$. □

The dominated convergence theorem is useful when the random variables are uniformly bounded by some constant K. In this particular case, the weaker convergence (in probability) can be assumed instead of the almost sure convergence. The following result will be used in the proof of Theorem 2.

Lemma 26 (Bounded convergence). *Assume that $X_n \xrightarrow{\mathbb{P}} X$ and that there exists a positive constant K such that almost surely, $\forall n, |X_n| \leq K$, then*

$$\mathbb{E}\left[|X_n - X|\right] \underset{n \to \infty}{\longrightarrow} 0.$$

Proof. The random variable X is also bounded in probability by K. Indeed, $|X| \leq |X - X_n| + |X_n| \leq |X - X_n| + K$, we have that $\mathbb{P}\left(|X| > K + \varepsilon\right) \leq \mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0$. Hence, $\mathbb{P}\left(|X| > K + \varepsilon\right) = 0, \forall \varepsilon > 0$ which means that $\mathbb{P}\left(|X| \leq K\right) = 1$. By conditioning,

$$\begin{aligned} \mathbb{E}\left[|X_n - X|\right] &= \mathbb{E}\left[|X_n - X| \mathbb{1}_{|X_n - X| > \varepsilon}\right] + \mathbb{E}\left[|X_n - X| \mathbb{1}_{|X_n - X| \leq \varepsilon}\right] \\ &\leq 2K\mathbb{P}\left(|X_n - X| > \varepsilon\right) + \varepsilon. \end{aligned}$$

□

13.5.2 Fatou Lemma

In the following, we denote by $a_n \uparrow a$, the simultaneity of $a_n \to a$ and a_n is increasing. (GIVE A GOOD LOCATION)

Lemma 27 (Fatou)**.** *For a sequence of non-negative measurable function $(f_n)_{n\in\mathbb{N}}$, we have that,*

$$\mu(\liminf f_n) \leq \liminf \mu(f_n).$$

A simple way to remember the order between $\int$ and $\liminf$, one of my teacher gave me the simple trick based on the lexical ordering : $il \leq li$ where l stands for the limit and i stands for the integral. This is interpreted as $\int \liminf \leq \liminf \int$.

Proof. Define the sequence $(g_k)_k$ by,

$$g_k = \inf_{n\geq k} f_n.$$

The sequence is well defined as a infimum of a sequence of non-negative numbers. By definition of $(g_k)_k$,

$$g_k \uparrow \liminf f_n,$$

and for any $n \geq k$, $f_n \geq g_k$, so that $\mu(f_n) \geq \mu(g_k)$ and then,

$$\mu(g_k) \leq \inf_{n\geq k} \mu(f_n).$$

Since (g_k) is non-decreasing, we can apply (MON) to get that

$$\mu(\liminf f_n) = \mu(\lim_k g_k) \overset{(MON)}{=} \lim_k \mu(g_k) \leq \lim_k \inf_{n\geq k} \mu(f_n) = \liminf \mu(f_n).$$

□

Lemma 28 (Reverse Fatou)**.** *Let $(f_n)_n$ be a sequence of measurable functions such that, for any n, $f_n \leq g$ with $\mu(g) < +\infty$, then*

$$\mu(\limsup f_n) \geq \limsup \mu(f_n)$$

Proof. Apply Fatou Lemma for $(g - f_n)_n$. □

13.6 The Monotone convergence theorem

13.6.1 Monotone convergence for measures

We begin with the monotone properties of measures. For measurable sets $(F_n)_n$ and F, the notation $F_n \uparrow F$ means $\forall n,\ F_n \subseteq F_{n+1}$ and $\bigcup F_n = F$ and $F_n \downarrow F$ means $\forall n,\ F_{n+1} \subseteq F_n$ and $\bigcap F_n = F$.

Lemma 29 (Monotone convergence for measures)**.** *Let $(\mathcal{X}, \mu)$ be a measure space, then*

1. *If $(F_n)_n$ are measurable sets such that $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$.*

2. *If $(G_n)_n$ are measurable sets such that $G_n \downarrow G$ and there exists k such that $\mu(G_k) < \infty$, then $\mu(G_n) \downarrow \mu(G)$.*

Proof. For 1., define $G_1 = F_1$ and $G_n := F_{n+1} \backslash F_n$ and remark that these are disjoints sets. As the measure of a countable union of disjoints sets equals the sum of the measures of the sets, we get

$$\mu(F_n) = \mu(\bigcup_{i=1}^{n} G_i) = \sum_{i=1}^{n} \mu(G_i) = \sum_{i=1}^{\infty} \mu(G_i) \uparrow \mu(F).$$

For 2., use 1. with $F_n = G_k \backslash G_{k+n}$, $F = G_k \backslash G$ and decompose $\mu(G_k) = \mu(G) + \mu(G_k \backslash G)$. □

13.6.2 Technical lemmas

Doubly monotone convergence

Lemma 30 (Doubly monotone sequences)**.** *Let $(a_{n,k})_{n\in\mathbb{N},k\in\mathbb{N}}$ be a double sequence of non-negative numbers. Assume that a is doubly monotone, which means*

1. $\forall k \in \mathbb{N},\ (a_{n,k})_n$ *is non-decreasing and* $\exists a_{\infty,k} \in [0,+\infty],\ a_{n,k} \underset{n\to\infty}{\longrightarrow} a_{\infty,k}$.
2. $\forall n \in \mathbb{N},\ (a_{n,k})_k$ *is non-decreasing and* $\exists a_{n,\infty} \in [0,+\infty],\ a_{n,k} \underset{k\to\infty}{\longrightarrow} a_{n,\infty}$.

Then,

$$\lim_k a_{\infty,k} = \lim_n a_{n,\infty}.$$

Proof. By a one-to-one transformation (by Arctan for example) of the sequence, we can assume it uniformly bounded. Let

$$a_\infty^{(1)} = \lim_k a_{\infty,k} \qquad \text{and} \qquad a_\infty^{(2)} = \lim_n a_{n,\infty}.$$

Now let $\varepsilon > 0$. Let k large enough, thus $n = n(k)$ large enough to get

$$a_{n,k} > a_{\infty,k} - \varepsilon > a_\infty^{(1)} - 2\varepsilon.$$

But $a_\infty^{(2)} \geq a_{n,\infty} \geq a_{n,k}$ which finally gives $a_\infty^{(2)} \geq a_\infty^{(1)}$. Repeating the argument symmetrically, we finally get the equality of the two limits. □

Staircase approximation

In the following result, we expose a way to define a sequence of simple functions increasingly converging to a given function.

Definition 17. *Let $\alpha_p : [0,+\infty] \to [0,+\infty]$ given by*

$$\alpha_p(x) = \begin{cases} 0 & \text{if } x = 0 \\ (i-1)2^{-p} & \text{if } (i-1)2^{-p} < x \leq i2^{-p} \leq p \ (\forall i \in \mathbb{N}) \\ p & \text{if } x > p \end{cases}$$

This function is left-continuous (i.e., if $x \to x_0$ with $x \leq x_0$, then $\alpha_p(x) \to \alpha_p(x_0)$).

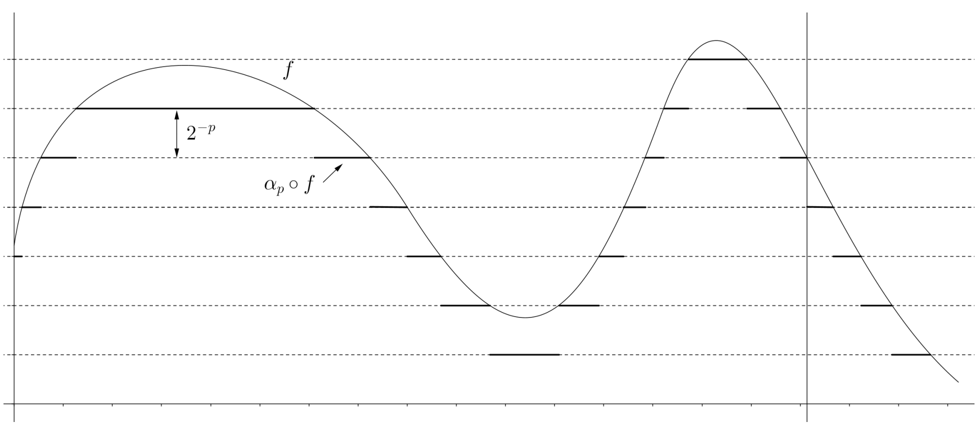


Figure 13.2: An example of the staircase transformation

Proposition 27. *The sequence $(\alpha_p \circ f)_p$ is a sequence of simple functions such that $\alpha_p \circ f \uparrow f$.*

A simpler case: Simple functions

Lemma 31. *Let $(f_n)_n$ be a sequence of non-negative simple functions and f a non-negative measurable function such that $f_n \uparrow f$, then*

$$\mu(f_n) \uparrow \mu(f).$$

Proof. Step 1 : ($f = \mathbb{1}_A$) Assume that $f_n \uparrow \mathbb{1}_A$, for A measurable. We obviously have that $\mu(f_n) \leq \mu(\mathbb{1}_A)$. Moreover, the sequence of real numbers $\mu(f_n)$ is non-decreasing. Let $\varepsilon > 0$ and $A_n = \{x \in A : f_n(x) > 1 - \varepsilon\}$. We have that $A_n \uparrow A$ and then, by Lemma 29, $\mu(A_n) \uparrow \mu(A)$. But, by definition,

$$(1-\varepsilon)\mathbb{1}_{A_n} \leq f_n$$

so that $(1-\varepsilon)\mu(A_n) \leq \mu(f_n)$. Since we took an arbitrary ε, it holds that

$$\mu(\mathbb{1}_A) = \mu(A) \leq \liminf \mu(f_n) \leq \limsup \mu(f_n) \leq \mu(\mathbb{1}_A)$$

Step 2 : (f a simple functions) Let f be of the form $f = \sum \alpha_k \mathbb{1}_{A_k}$, for a finite number of A_k. We apply the previous case to the convergence of

$$\alpha_k^{-1}\mathbb{1}_{A_k} f_n \uparrow \mathbb{1}_{A_k}$$

Step 3 : (Approximating f) We show that there exists a sequence f_k of simple functions satisfying both $\mu(f_k) \uparrow \mu(f)$ and $f_k \uparrow f$. By definition of the Lebesgue integral,

$$\mu(f) = \sup\{\mu(h) : \ h \text{ is simple and } 0 \leq h \leq f\}.$$

Hence, there exists a sequence (h_k) such that $\mu(h_k) \uparrow \mu(f)$. But using the staircase function α_p, we can construct a sequence $g_p := \alpha_p \circ f$ such that $g_p \uparrow f$. Now define

$$\overline{f}_k = \max\{g_k, h_1, \ldots, h_k\}.$$

Since $(g_k)_k$ is non-decreasing, $\overline{f}_k$ is also non-decreasing and $\mu(h_k) \leq \mu(\overline{f}_k) \leq \mu(f)$ and so holds the convergence $\mu(f_k) \to \mu(f)$.

Step 4 : (Uniqueness of the limit) Let $f_n \uparrow f$ and $g_k \uparrow f$ two non-decreasing sequences of simple functions. We show that $\lim \mu(f_n) = \lim \mu(g_k)$. Define $h_{n,k} = \min\{f_n, g_k\}$ and note that it is a doubly increasing sequence. Moreover,

$$h_{n,k} \underset{n\to\infty}{\longrightarrow} g_k \qquad \text{and} \qquad h_{n,k} \underset{k\to\infty}{\longrightarrow} f_n.$$

Since the limits g_k, f_n and $h_{n,k}$ are simples functions, we can apply Step 2 and get

$$\mu(h_{n,k}) \underset{n\to\infty}{\longrightarrow} \mu(g_k) \qquad \text{and} \qquad \mu(h_{n,k}) \underset{k\to\infty}{\longrightarrow} \mu(f_n)$$

which allows us to apply Lemma 30 to the sequence $\mu(h_{n,k})_{n,k}$ and we get the uniqueness of the limit.

Step 5 : (Putting all together) Take $\overline{f}_k$ defined in step 3, then $\mu(\overline{f}_k) \uparrow \mu(f)$. But, by hypothesis, we have that $f_n \uparrow f$, then by the uniqueness of the limit $\mu(f_n) \uparrow \mu(f) = \lim \mu(\overline{f}_k)$. □

Monotone convergence theorem

Theorem 15 ((MON))**.** *Let $(f_n)_n$ and f non-negative measurable functions such that $f_n \uparrow f$. Then*

$$\mu(f_n) \uparrow \mu(f).$$

Proof. By the staircase approximation, we construct a double index sequence $(\alpha_p \circ f_n)_{n,p}$ of simple functions such that

$$\alpha_p \circ f_n \underset{p\to\infty}{\longrightarrow} f_n \qquad \text{and} \qquad \alpha_p \circ f_n \underset{p\to\infty}{\longrightarrow} \alpha_p \circ f$$

where the first fact holds by the definition of α_p and the second holds by the left-continuous property of α_p. Obviously, the convergences occur in an increasing manner. Then applying Lemma 31, we get

$$\mu(\alpha_p \circ f_n) \underset{p\to\infty}{\longrightarrow} \mu(f_n) \qquad \text{and} \qquad \mu(\alpha_p \circ f_n) \underset{p\to\infty}{\longrightarrow} \mu(\alpha_p \circ f)$$

which occurs again in an increasing manner. Now applying Lemma 30 for the sequence $(\mu(\alpha_p \circ f_n))_{n,p}$, we get

$$\mu(f_n) \uparrow \lim_{p\to+\infty} \mu(\alpha_p \circ f) = \mu(f).$$

□

Chapter 14

Basic probability results

We state here the important Borel-Cantelli lemma.
For a sequence of events $(E_n)_n$ we denote $\{E_n \text{ i.o.}\}$ for the event

$$\begin{aligned}\{E_n \text{ i.o.}\} &= \{\omega : \forall m, \exists n(\omega) \geq m \text{ such that } \omega \in E_{n(\omega)}\}.\\ &= \{\omega : \omega \in E_n \text{ for infinitely many } n\}\end{aligned}$$

Lemma 32 (Borel-Cantelli)**.** *For a sequence of events $(E_n)_n$ such that $\sum_{n\geq 0} \mathbb{P}(E_n) < +\infty$. Then*

$$\mathbb{P}(\limsup E_n) = \mathbb{P}(E_n \text{ } i.o.) = 0$$

Proof. Defining $G_m := \bigcup_{n\geq m} E_n$ and $G := \limsup E_n$ so that we have $G_m \downarrow G$. Then for any $m \in \mathbb{N}$, we have

$$\mathbb{P}(G) \underset{\text{Lemma 29}}{\leq} \mathbb{P}(G_m) \leq \sum_{n\geq m} \mathbb{P}(E_n).$$

When we let $m \to +\infty$, $\sum_{n\geq m} \mathbb{P}(E_n) \underset{n\to\infty}{\longrightarrow} 0$ and then $\mathbb{P}(G) = 0$. □

Lemma 33 (Borel-Cantelli-reverse)**.** *For a sequence of independent events $(E_n)_n$ such that $\sum_{n\geq 0} \mathbb{P}(E_n) = +\infty$ one has*

$$\mathbb{P}(\limsup E_n) = \mathbb{P}(E_n \text{ } i.o.) = 1$$

Proof. We work with the complementary of E_n and we also note $p_n = \mathbb{P}(E_n)$. By independence,

$$\mathbb{P}\left(\bigcap_{n\geq m} E_n^c\right) = \prod_{n\geq m} (1 - p_n), \quad \forall m,$$

where this can be shown for every $r \geq n \geq m$ to get a finite intersection first and then let $r \to \infty$. But since $1 - x \leq e^{-x}$ for $x \geq 0$, one has that

$$\prod_{n\geq m} (1 - p_n) \leq \exp\left(-\sum_{n\geq m} p_n\right) = 0.$$

But since $(\limsup E_n)^c = \liminf E_n^c = \bigcup_m \bigcap_{n\geq m} E_n^c$, we get that $\mathbb{P}((\limsup E_n)^c) = 0$. □

Lemma 34 (Jensen Inequality)**.** *Let ϕ be a convex function on an open interval I of $\mathbb{R}$ of the form (a, b). For a random variable X such that*

$$\mathbb{E}[|X|] < +\infty, \qquad \mathbb{P}(X \in I) = 1, \qquad \mathbb{E}[|\phi(X)|] < +\infty.$$

Then we have that

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Proof. Let $(a_n)_n$ and $(b_n)_n$ defined in Lemma 22, in order to have $\phi(x) = \sup_{n\in\mathbb{N}}(a_n x + b_n)$. Then, for any n,

$$\mathbb{E}[\phi(X)] \geq a_n \mathbb{E}[X] + b_n.$$

But since the inequality is valid for all n, the $\sup_n$ is also bounded by $\mathbb{E}[\phi(X)]$ which gives the result. □

14.0.1 Convergence in probability

The following results are stated for random variables taking values in $\mathbb{R}$. At the simple cost of replacing $|X-Y|$ by the quantity $d(X,Y)$ defined in Definition 2, we can generalize the following results to random vectors in $\mathbb{R}^k$.

Lemma 35. *Let $(X_n)_n$ be a sequence of random variables such that*

$$\forall \varepsilon > 0, \ \sum_{n=0}^{+\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) < +\infty$$

then $X_n \xrightarrow{a.s.} X$.

Proof. Let $E_{n,\varepsilon} := \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}$ and let $A_\varepsilon := \limsup E_{n,\varepsilon}$. The assumption of Borel-Cantelli lemma is fulfilled and thus $\mathbb{P}(A_\varepsilon) = 0$. But

$$A_\varepsilon^c = \{\omega \in \Omega : \ \exists n_0, \forall n \geq n_0, \ |X_n(\omega) - X(\omega)| < \varepsilon\}$$

is then of probability 1. Let $\varepsilon_i = 2^{-i}$ and let

$$\Lambda := \bigcap_{i=0}^{\infty} A_{\varepsilon_i}^c.$$

The set Λ is a countable intersection of events of probability one then is also of probability 1. Now for any $\omega \in \Lambda$, we have that $X_n(\omega) \to X(\omega)$. This is exactly $X_n \xrightarrow{a.s.} X$. □

We see directly that the assumption of Lemma 35 implies the convergence in probability of the sequence X_n towards X. The convergence of probability does not implies convergence almost sure as seen by Example 2.

Lemma 36. *Let $(X_n)_n$ be a sequence of random variables such that $X_n \xrightarrow{\mathbb{P}} X$. Then there exists a sub-sequence $(X_{n_k})_k$ such that $X_{n_k} \xrightarrow{a.s.} X$.*

Proof. We will extract a sub-sequence of the sequence $(X_n)_n$ which verifies the assumption of Lemma 35. Let $\varepsilon_k = 2^{-k}$. The convergence in probability implies that $\mathbb{P}(|X_n - X| > \varepsilon_k) \underset{n\to\infty}{\longrightarrow} 0$ then $\exists n_k \in \mathbb{N}$ such that

$$\mathbb{P}\left(|X_{n_k} - X| > \varepsilon_k\right) \leq \frac{1}{k^2}.$$

Let $\varepsilon > 0$. There exists $k_0 \in \mathbb{N}$ such that $\forall k \geq k_0$, $\varepsilon_k < \varepsilon$, then

$$\{|X_{n_k} - X| > \varepsilon\} \subset \{|X_{n_k} - X| > \varepsilon_k\}.$$

We verify the assumption of Lemma 35,

$$\sum_{k=0}^{+\infty} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right) \leq \underbrace{\sum_{k=0}^{k_0-1} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right)}_{<+\infty} + \sum_{k=k_0}^{+\infty} \underbrace{\mathbb{P}\left(|X_{n_k} - X| > \varepsilon_k\right)}_{\text{summable}} < +\infty$$

and then $X_{n_k} - X \xrightarrow{a.s.} 0$. □

14.0.2 From convergence in $\mathbb{P}$ to a.s.

In this section, we give a simple argument that permits to bridge the gap between convergence in probability and convergence a.s. This is doable when the random variables are upper bounded by a common variable.

Lemma 37 (Kolmogorov Truncation)**.** *Let $X_1, \ldots, X_n, \ldots$ be random vectors such that there exists X a positive random variable with $\mathbb{E}[X] < \infty$ and $\forall n \in \mathbb{N}^*$, $\|X_n\| \leq X$. For all $n \in \mathbb{N}^*$, define*

$$Y_n := \begin{cases} X_n & \text{if } \|X_n\| \leq n \\ 0 & \text{if } \|X_n\| > n \end{cases}$$

Then,

i) $\mathbb{P}(X_n = Y_n$ eventually$) = 1$. [PRECISE THIS]

ii) $\|\sum_{n\geq 1} n^{-2}\, Var(Y_n)\| < \infty$.

Proof. For proving i), we use Borel-Cantelli's lemma (Lemma 32) and the fact that

$$\sum_{n\geq 1}\mathbb{P}\left(Y_n\neq X_n\right)=\sum_{n\geq 1}\mathbb{P}\left(\|X_n\|>n\right)\leq\sum_{n\geq 1}\mathbb{P}\left(X>n\right)\leq\mathbb{E}\left[X\right]<\infty.$$

For ii), we see that

$$\begin{aligned}\|\sum_{n\geq 1}n^{-2}\mathrm{Var}\left(Y_n\right)\| &\leq\sum_{n\geq 1}n^{-2}\mathbb{E}\left[\|Y_n\|^2\right]\leq\sum_{n\geq 1}\frac{\mathbb{E}\left[\|X_n\|^2\mathbb{1}_{\|X_n\|\leq n}\right]}{n^2}\leq\sum_{n\geq 1}\frac{\mathbb{E}\left[\|X_n\|^2\mathbb{1}_{\|X_n\|\leq n}\mathbb{1}_{X\leq n}\right]}{n^2}+\sum_{n\geq 1}\mathbb{E}\left[\mathbb{1}_{X>n}\right]\\ &\leq\sum_{n\geq 1}\frac{\mathbb{E}\left[X^2\mathbb{1}_{X\leq n}\right]}{n^2}+\mathbb{E}\left[X\right]=\mathbb{E}\left[X^2\sum_{n\geq\max(1,X)}\frac{1}{n^2}\right]+\mathbb{E}\left[X\right]\\ &\leq 2\mathbb{E}\left[X^2\sum_{n\geq\max(1,X)}\frac{1}{n}-\frac{1}{n+1}\right]+\mathbb{E}\left[X\right]=2\mathbb{E}\left[\frac{X^2}{\max(1,X)}\right]+\mathbb{E}\left[X\right]\leq 3\mathbb{E}\left[X\right]<\infty\end{aligned}$$

□

This later result allows to derive a implication between convergence in probability and convergence a.s. for sums of random variables.

Lemma 38. *Let $X_1,\dots,X_n,\dots$ be random vectors such that there exists X a positive random variable with $\mathbb{E}\left[X\right]<\infty$ and $\forall n\in\mathbb{N}^*,\ \|X_n\|\leq X$. We assume that*

$$S_n=\frac{1}{n}\sum_{i=1}^{n}X_n\xrightarrow{\mathbb{P}}\mu.$$

Then,

$$S_n\xrightarrow{a.s.}\mu.$$

Proof. Since the sequence $(X_i)_i$ is uniformly bounded by X which is integrable, we have that it is U.I. (see Proposition 1) and so is $(S_n)_n$. Hence, one has that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_n\right]\underset{n\to+\infty}{\longrightarrow}\mu.$$

Now, using the Y_i of Lemma 37, we get that

$$\frac{1}{n}\sum_{i=1}^{n}Y_i-\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{a.s.}0\quad\text{and also}\quad\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_n\right]\underset{n\to+\infty}{\longrightarrow}\mu\text{ (by DOM)}.$$

Then, it only remains to show that $n^{-1}\sum Y_i-\mathbb{E}\left[Y_i\right]\xrightarrow{a.s.}0$. The second point of Lemma 37 allows us to use Lemma 35 together with Bienaymé-Chebyshev inequality to get the conclusion. □

Remark 4. *Notice that the same trick can be used to show that*

$$\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}X_{i,t}\xrightarrow{\mathbb{P}}0\quad\Leftrightarrow\quad\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}X_{i,t}\xrightarrow{a.s.}0$$

under the uniform assumption $\forall i,\ \|X_{i,t}\|\leq X_t$ such that $\mathbb{E}\left[\sup X_t\right]<\infty$. In this case, point i) of Lemma 37 will be replaced by $\mathbb{P}\left(\forall t\in\mathcal{T},X_{i,t}=Y_{i,t}\text{ eventually}\right)=1$.

Exercice 20. *Show the equivalence of Remark 4.*

Chapter 15

Carathéodory theorem

15.1 Measure set theory

15.1.1 Special class of sets

Algebras For a set Ω, we define an **algebra** as a collection Σ_0 of subsets of Ω such that

- $\Omega \in \Sigma_0$.
- If $F \in \Sigma_0$ then $F^c \in \Sigma_0$. (*Stable under complementation*)
- If $F_1, F_2 \in \Sigma_0$ then $F_1 \cup F_2 \in \Sigma_0$. (*Stable under finite union*).

σ-algebras A collection Σ of subsets of Ω is a **σ-algebra** if

- Σ is an algebra.
- $F_1, F_2, \ldots, F_n, \cdots \in \Sigma$ then $\bigcup_{n\in\mathbb{N}} F_n \in \Sigma$. (*Stable under countable union*)

In the context of σ-algebras, we omit the index 0 in the notation of Σ. This is to strengthen the fact that σ-algebras are the main purpose of measure theory.

Comments 1. *Note that it is always possible to assume that the sequence of elements are disjoints since, one may replace the sequence by $G_1 = F_1, G_2 = F_2 \backslash F_1, \ldots, G_n = F_n \backslash \bigcup_{i=1}^{n-1} F_i, \ldots$ which is such that*

$$\bigcup_{n\in\mathbb{N}} F_n = \bigcup_{n\in\mathbb{N}} G_n.$$

π-systems A collection Σ_0 of subsets of Ω is a **π-system** if

- $F_1, F_2 \in \Sigma_0$ then $F_1 \cap F_2 \in \Sigma_0$. (*Stable under finite intersection*)

It is direct to see that any σ-algebra is an algebra and any algebra is a π-system.

λ-sets For a function $\lambda : \Sigma_0 \to [0, +\infty]$ on the algebra Σ_0 and such that $\lambda(\emptyset) = 0$, we say that a element $L \in \Sigma_0$ is a **λ-set** if

$$\forall K \in \Sigma_0,\ \lambda(L \cap K) + \lambda(L^c \cap K) = \lambda(K). \tag{15.1}$$

σ-algebras generated For a class $\mathcal{C}$ of subsets of Ω, we define the **σ-algebra generated** by $\mathcal{C}$ and denoted by $\sigma(\mathcal{C})$ as the smallest (for the inclusion) σ-algebra that contains $\mathcal{C}$. In more precise words, $\sigma(\mathcal{C})$ is the intersection (show that it is still a σ-algebra) of all σ-algebras that contain $\mathcal{C}$.

15.1.2 Definition of measures

As in the previous section, we define special classes of functions $\Sigma_0 \to [0, +\infty]$ adapted to each context of subsets defined above.

Additivity Let Σ_0 be an algebra. A function $\mu_0 : \Sigma_0 \to [0, +\infty]$ is said to be **finitely additive** (or additive) if

- $\mu_0(\emptyset)$=0.
- For any pair of *disjoints* sets $F_1, F_2 \in \Sigma_0$, we have

$$\mu_0(F_1 \cup F_2) = \mu_0(F_1) + \mu_0(F_2).$$

Measure Let Σ be an σ-algebra. A function $\mu : \Sigma \to [0, +\infty]$ is said to be **a measure** (or countably additive) if

- $\mu(\emptyset)$=0.
- For any sequence of *disjoints* sets $F_1, F_2, \dots, F_n, \dots \in \Sigma$, we have

$$\mu(\bigcup_{n\in\mathbb{N}} F_n) = \sum_{n\in\mathbb{N}} \mu(F_n).$$

All together the triple Ω, Σ, μ is called a measure space. The measure μ is said to be **finite** if $\mu(\Omega) < +\infty$. *mu* is said to be **σ-finite** if there exists a sequence $\Omega_1, \dots, \Omega_n, \dots$ of elements of Σ such that

$$\bigcup_{n\in\mathbb{N}} \Omega_n = \Omega \quad \text{and} \quad \mu(\Omega_n) < +\infty, \forall n \in \mathbb{N}.$$

A **probability space** is a measure space Ω, Σ, μ where $\mu(\Omega) = 1$ and the measure μ is called a **probability measure**. We usually adopt the notation P instead of μ for a probability measure.
A more general notion of measure is the so-called outer measures that are a building step to construct important examples of measures such that Lebesgue measure.

Outer measures Let Σ be a σ-algebra. A function $\mu_0 : \Sigma \to [0, +\infty]$ is called a **outer measure** if it satisfies

- $\mu_0(\emptyset) = 0$.
- (increasing) For any two sets $F_1, F_2 \in \Sigma$ such that $F_1 \subseteq F_2$,

$$\mu_0(F_1) \leq \mu_0(F_2).$$

- (countable sub-additivity) For any sequence $F_1, \dots, F_n, \dots$ of elements of Σ,

$$\mu_0(\bigcup_{n\in\mathbb{N}} F_n) \leq \sum_{n\in\mathbb{N}} \mu_0(F_n).$$

15.1.3 Extension theorems

Proposition 28 (λ-sets form an algebra)**.** *Let $\mathcal{L}_0$ be the set of all λ-sets of an algebra Σ_0. Then the set $\mathcal{L}_0$ is an algebra and the restriction $\lambda_{\mathcal{L}_0} : \mathcal{L}_0 \to [0, +\infty]$ is additive.*

Proof. We verify the three axioms of an algebra.
Full set Ω is obviously a λ-set.
Complementary By the symmetry of the definition of a λ-set, its complementary is trivially a λ-set.
Stability by finite intersection Let L_1 and L_2 two elements of $\mathcal{L}_0$, let $L = L_1 \cap L_2$ and let $K \in \Sigma_0$. Since L_1, L_2 are λ-sets, we get that

$$\begin{aligned}
\lambda(L \cap K) + \lambda(L_1^c \cap L_2 \cap K) &= \lambda(L_2 \cap K) && \text{(with } L_1 \text{ and } L_2 \cap K)\\
\lambda(L_2 \cap K) + \lambda(L_2^c \cap K) &= \lambda(K) && \text{(with } L_2 \text{ and } K)\\
\lambda(L^c \cap K) &= \lambda(L_2 \cap L_1^c \cap K) + \lambda(L_2^c \cap K) && \text{(with } L_2 \text{ and } L^c \cap K)
\end{aligned}$$

where we remark that $L^c \cap L_2 = L_2 \cap L_1^c$ and $L^c \cap L_2^c = L_2^c$. Now summing up the three equalities leads to the desired equation for L.
λ is finitely additive Let L_1 and L_2 two disjoints λ-sets. Using Equation (15.1) for L_1 and $K = L_1 \cup L_2$, we get

$$\lambda(L_1 \cup L_2) = \lambda((L_1 \cup L_2) \cap L_1) + \lambda((L_1 \cup L_2) \cap L_1^c) = \lambda(L_1) + \lambda(L_2)$$

which finishes the proof. □

The following lemma explores the case of σ-algebras instead of simple algebras. Its stronger structure permits to deduce that μ_0 is a measure at the cost of assuming that it is already a outer measure.

Lemma 39 (Carathéodory Lemma)**.** *Let λ be a outer measure on (Ω, Σ). The class $\mathcal{L}$ of all the λ-sets in Σ is a σ-algebra on which the outer measure λ is a measure.*

Proof. Thanks to the result of Proposition 28, we already know that λ is additive. Hence, the only two things that remains to show is the countable additivity for μ_0 and the stability under countable union for $\mathcal{L}$. Let $L_1, \ldots, L_n, \ldots$ be a sequence of disjoints elements in $\mathcal{L}$. Let $L = \bigcup_{n\geq 1} L_n$. By the fact that any finite union of elements in $\mathcal{L}$ is again in $\mathcal{L}$, we get that for $M_n = \bigcup_{k=1}^n L_k$ and any $K \in \Sigma$,

$$\lambda(K) = \lambda(M_n \cap K) + \lambda(M_n^c \cap K) \geq \lambda(M_n \cap K) + \lambda(L^c \cap K)$$

since $L^c \subseteq M_n^c$. But then, using Proposition 28 again leads to the following inequality

$$\lambda(K) \geq \sum_{k=1}^{n} \lambda(L_k \cap K) + \lambda(L^c \cap K) \quad (\forall n \geq 1),$$

and taking the limit and the countable sub-additivity we finally get

$$\lambda(K) \geq \sum_{k\geq 1} \lambda(L_k \cap K) + \lambda(L^c \cap K) \geq \lambda(L \cap K) + \lambda(L^c \cap K).$$

On the other side, the sub-additivity of λ implies,

$$\lambda(K) \leq \lambda(L \cap K) + \lambda(L^c \cap K)$$

and then the two previous inequalities imply that all the inequalities written above are actual equalities. In particular, this shows that L belongs to $\mathcal{L}$ (and then $\mathcal{L}$ is a σ-algebra) and taking $K = L$ we see that

$$\lambda(L) = \sum_{k\geq 1} \lambda(L_k).$$

□

15.1.4 Carathéodory theorem

The following theorem is an angular stone to construct all the measures that are commonly used in probabilistic theory.

Theorem 16. *Let Ω be a set, and let Σ_0 be an algebra on Ω. We associate to Σ_0 its generated σ-algebra $\Sigma = \sigma(\Sigma_0)$. Let μ_0 be a countably sub-additive map $\mu_0 : \Sigma_0 \to [0, +\infty]$.* ***Then****, there exists a measure $\mu : \Sigma \to [0, +\infty]$ such that*

$$\mu_{|\Sigma_0} = \mu_0.$$

Moreover, if $\mu_0(\Omega) < +\infty$, then the extension μ is unique.

Remark Many authors do assume that the map μ_0 is countably additive in Theorem 16. It is actually not needed as seen in the proof below. Besides, it is usually of similar complexity to show countable sub-additivity or countable additivity. As a corollary result, we get that μ_0 is in fact countable additive as a restriction of μ.

Proof. We consider the largest σ-algebra possible $\mathcal{G}$ that contain all the subsets of Ω. We define a function $\lambda : \mathcal{G} \to [0, +\infty]$ by

$$\lambda(G) = \inf \sum_{n\geq 1} \mu_0(F_n) \quad (\forall G \in \mathcal{G})$$

where the infimum is taken over all the sequences $(F_n)_n$ of elements of Σ_0 such that $G \subseteq \bigcup_{n\geq 1} F_n$.

Fact 1 : λ is an outer measure on $(\Omega, \mathcal{G})$

It is direct to see that $\lambda(\emptyset) = 0$. It is also direct to get the increasing property since the definition of λ involves an inf. For the sub-additivity, let $(G_n)_n$ be a sequence of elements of $\mathcal{G}$ such that $\lambda(G_n) < +\infty$ (otherwise there is nothing to prove). Then, for any $n \geq 1$ and $\varepsilon > 0$, it is possible to find a sequence $(F_{n,k})$ of elements of Σ_0 such that

$$G_n \subseteq \bigcup_{k\geq 1} F_{n,k} \quad \text{and} \quad \sum_{k\geq 1} \mu_0(F_{n,k}) < \lambda(G_n) + \varepsilon 2^{-n}.$$

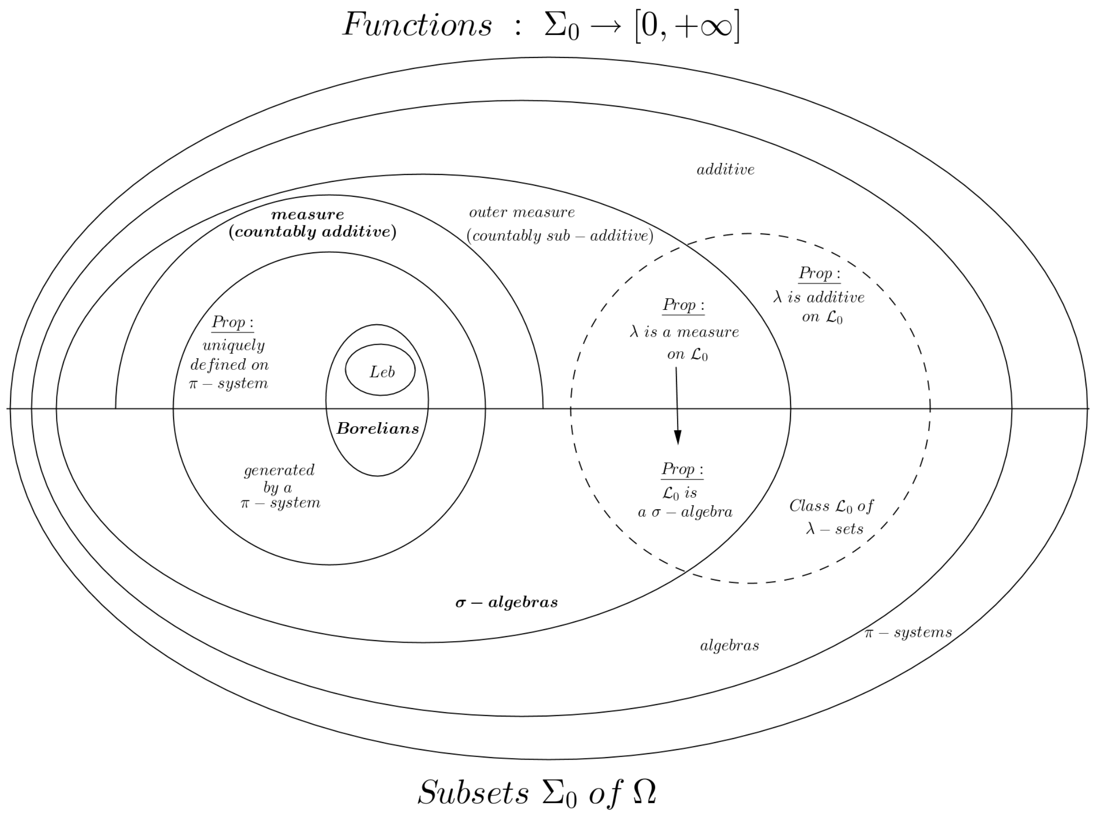


Figure 15.1: A sum up of the classes of importance in measure theory represented as inclusion of sets for sub-classes. On the bottom side, the definitions of different types of classes correspond to definitions for non-negative valued function on the top. The inclusions represents sub-classes and bold notions are enlightened to show their major importance. Finally, dashed lines are reserved for minor notions.

Let $G = \bigcup_{n\geq 1} G_n \subseteq \bigcup_{n,k\geq 1} F_{n,k}$ so that $(F_{n,k})_{n,k}$ is a sequence of elements of Σ_0 containing G. Then,

$$\lambda(G) \leq \sum_{n,k\geq 1} \mu_0(F_{n,k}) < \sum_{n\geq 1} \lambda(G_n) + \varepsilon$$

and since, ε is arbitrary, we get the sub-additivity.

Fact 2 : λ is a measure on $(\Omega, \mathcal{L})$
We define $\mathcal{L}$ the class of λ-sets on the class $\mathcal{G}$. By Carathéodory Lemma 39, we get that $\mathcal{L}$ is a σ-algebra and λ is indeed a measure on $\mathcal{L}$.

Fact 3 : $\lambda = \mu_0$ on (Ω, Σ_0)
Let $F \in \Sigma_0$. We have directly that $\lambda(F) \leq \mu_0(F)$ (pick a silly sequence). For the $\lambda(F) \geq \mu_0(F)$ part, pick any sequence $(F_n)_n$ of elements of Σ_0 with an union containing F and define the sequence of disjoints sets $(E_n)_n$, by

$$E_1 := F_1, \qquad E_n = F_n \backslash (\bigcup_{k=1}^{n-1} F_k).$$

Then, by the countable sub-additivity of μ_0, we get

$$\mu_0(F) = \mu_0(\bigcup_{n\geq 1}(F \cap E_n)) \leq \sum_{n\geq 1} \mu_0(F \cap E_n) \leq \sum_{n\geq 1} \mu_0(E_n) \leq \sum_{n\geq 1} \mu_0(F_n).$$

Now, taking the infimum on both sides gives $\mu_0(F) \leq \lambda(F)$ hence the equality.

Fact 4 : $\Sigma_0 \subseteq \mathcal{L}$
Let $F \in \Sigma_0$ and $K \in \mathcal{G}$. We will show that F is a λ-set. By the sub-additivity of λ, we already have that

$$\lambda(K) \leq \lambda(F \cap K) + \lambda(F^c \cap K).$$

For any $\varepsilon > 0$, there exists a sequence $(F_n)_n$ of elements of Σ_0 such that $K \subseteq \bigcup_{n\geq 1} F_n$ and

$$\sum_{n\geq 1} \mu_0(F_n) < \lambda(K) + \varepsilon.$$

But, we also have

$$\sum_{n\geq 1} \mu_0(F_n) = \sum_{n\geq 1} \mu_0(F \cap F_n) + \sum_{n\geq 1} \mu_0(F^c \cap F_n) \geq \lambda(F \cap K) + \lambda(F^c \cap K).$$

Since, ε is arbitrary, we get that $\lambda(K) \geq \lambda(F \cap K) + \lambda(F^c \cap K)$ which concludes the fact.

Fact 5 : Definition of μ
By the fact 2,3 and 4, we get that $\Sigma_0 \subseteq \Sigma := \sigma(\Sigma_0) \subseteq \mathcal{L}$. But since we already defined λ, a measure extending μ_0 on $\mathcal{L}$, it suffices to define μ as the restriction of λ on Σ.
Fact 6 : Uniqueness of μ
In the case of $\mu(\Omega) < \infty$, we use Theorem 17 to conclude. □

A important side result of the proof that we gave here is a general construction of an outer measure on any algebra.

Canonical outer measure To any algebra Σ_0 defined on Ω, one can construct an outer measure by the formula

$$\lambda(G) = \inf \sum_{n\geq 1} \mu_0(F_n) \quad (\forall G \in \mathcal{P}(\Omega)) \tag{15.2}$$

where the infimum is taken over all the sequences $(F_n)_n$ of elements of Σ_0 such that $G \subseteq \bigcup_{n\geq 1} F_n$. Such an outer measure is named the **canonical outer measure** associated to Σ_0. But one has to be careful since a little structure (namely the sub-additivity) on μ_0 is needed to have that λ and μ_0 coincide on Σ_0.

15.1.5 Uniqueness of extension

In this section, we treat the case of the uniqueness of the extension of measures. In fact, it is sufficient to define the values of the measure on a smaller set than the σ-algebra Σ. The adapted notion is the π-systems. From the definitions, it is clear that σ-algebras are a stronger structure than π-systems. What is lacking from a π-system to be a σ-algebra is precisely the topic of d-systems (for Dynkin) defined in the following.

d-systems Let Ω be a set and $\mathcal{D}$ be a collection of subsets of Ω having the three following properties:

- $\Omega \in \mathcal{D}$.
- For any two elements $A, B \in \mathcal{D}$ with $A \subseteq B$, we have $B \backslash A \in \mathcal{D}$.
- For any sequence $(A_n)_n$ of elements of $\mathcal{D}$ such that $A_n \uparrow A$, then $A \in \mathcal{D}$.

Such a set $\mathcal{D}$ is called a **d-system**. For a class of subsets Σ_0, we denote by $d(\Sigma_0)$ the **generated d-system** as the set given by the intersection of all d-systems containing Σ_0.

Proposition 29. *Let Σ be a class of subsets of Ω. Then Σ is a σ-algebra if and only if it is a π-system and a d-system.*

Proof. We only need to prove the if part since, obviously, a σ-algebra is a π-system and a d-system. Assume that Σ is a π-system and d-system. If $F \in \Sigma$, then $F^c = \Omega \backslash F \in \Sigma$. Also for $F_1, F_2 \in \Sigma$, we have $F_1^c \cap F_2^c \in \Sigma$ (π-system) and $F_1 \cup F_2 = \Omega \backslash (F_1^c \cap F_2^c) \in \Sigma$, so that Σ is an algebra. Now let $(F_n)_n$ be a sequence in Σ and $G_n = F_1 \cup \cdots \cup F_n$. Obviously, $G_n \uparrow \bigcup F_k$ and then $\bigcup F_k \in \Sigma$. □

It is now the time to give the important result of the section.

Lemma 40 (Dynkin)**.** *Let Σ_0 be a π-system. Then*

$$d(\Sigma_0) = \sigma(\Sigma_0).$$

Proof. It is obvious that we have $d(\Sigma_0) \subseteq \sigma(\Sigma_0)$ so it is enough to show that $d(\Sigma_0)$ is a π-system. For that purpose, define

$$\mathcal{D}_1 := \{A \in d(\Sigma_0) : \ \forall B \in \Sigma_0, \ A \cap B \in d(\Sigma_0)\}$$

and

$$\mathcal{D}_2 := \{A \in d(\Sigma_0) : \ \forall B \in d(\Sigma_0), \ A \cap B \in d(\Sigma_0)\}.$$

We have $\mathcal{D}_2 \subseteq \mathcal{D}_1 \subseteq d(\Sigma_0)$ and we will show equality of these sets. First, we see that $\Sigma_0 \subseteq \mathcal{D}_1$ (since Σ_0 is a π-system). Thus, it is enough to show that $\mathcal{D}_1$ is a d-system. To see that, write for $A_1 \subseteq A_2$ two elements of $d(\Sigma_0)$ and $B \in \Sigma_0$,

$$(A_2 \backslash A_1) \cap B = (A_2 \cap B) \backslash (A_1 \cap B)$$

and for a sequence $A_n \uparrow A$ in $d(\Sigma_0)$,

$$(A_n \cap B) \uparrow (A \cap B).$$

The set $\mathcal{D}_1$ being a d-system, we have that $\mathcal{D}_1 = d(\Sigma_0)$. By definition of $\mathcal{D}_1$, this last fact insures that $\Sigma_0 \subseteq \mathcal{D}_2$. But as before, $\mathcal{D}_2$ is actually a d-system then $\mathcal{D}_2 = \Sigma_0$ and this shows that $d(\Sigma_0)$ is a π-system then a σ-algebra. Finally, $d(\Sigma_0) = \sigma(\Sigma_0)$. □

We are now ready to prove the following uniqueness result.

Theorem 17 (Uniqueness of extension)**.** *Let Ω be a set such that Σ_0 is a π-system on Ω. We define $\Sigma = \sigma(\Sigma_0)$. Let μ_1 and μ_2 be two measures on (Ω, Σ) such that*

- $\mu_1(\Omega) = \mu_2(\Omega) < \infty$.
- $\forall A \in \Sigma_0,\ \mu_1(A) = \mu_2(A)$.

Then,

$$\mu_1 = \mu_2 \quad \textit{as measures on } (\Omega, \Sigma).$$

Proof. Let $\mathcal{D} := \{A \in \Sigma : \ \mu_1(A) = \mu_2(A)\}$. The goal is to show that $\mathcal{D}$ is a d-system. For any $A, B \in \mathcal{D}$ with $A \subseteq B$, we have that

$$\mu_1(B \backslash A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \backslash A)$$

where the equality holds since we are only dealing with finite values. Then $B \backslash A \in \mathcal{D}$. Let $A_n \uparrow A$ where $A_n \in \mathcal{D}$, then

$$\mu_1(A) = \uparrow \lim \mu_1(A_n) = \uparrow \lim \mu_2(A_n) = \mu_2(A)$$

where we used Lemma 29. Thus $A \in \mathcal{D}$ and $\mathcal{D}$ is a d-system. We have $\Sigma_0 \subseteq \mathcal{D}$ then, using Dynkin's Lemma, we get that $\mathcal{D} = \Sigma$. □

Remarks The assumption on the finiteness of $\mu(\Omega)$ is important and cannot be avoided. The assumption that μ_1 and μ_2 are two measures is also important to use Lemma 29. The conclusion also fails to hold if μ_1 and μ_2 are only assumed to be finitely additive.

15.1.6 Definiton of the Lebesgue measure

The construction of Lebesgue measure is an important step to understand the classical construction of Skorokod for the existence of random variables of given distribution function. There is actually two options to define a measure based on a restriction of outer measures. The first one is to use Carathéodory extension theorem directly and then the only thing to check is the *sub-additivity* of μ_0. The second is to define the outer measure directly and to show that the outer measure defined in Equation (15.2) equals μ_0 on the algebra. We follow the second option here. The interested reader may find the other option in [15, A.1.9].

Definition of Leb **on** $((0,1], \mathcal{B}((0,1]))$

We define an algebra,

$$\Sigma_0 := \{A = (a_1, b_1] \cup \cdots \cup (a_r, b_r] : \ r \geq 1,\ a_i \leq b_i \leq a_{i+1} \leq b_{i+1}, \forall i\}.$$

as the set of all finite disjoint unions of semi-open intervals. It is easy to see that $\sigma(\Sigma_0) = \mathcal{B}((0,1])$. We can easily define a countably additive map μ_0 on Σ_0 by

$$\mu_0(A) := \sum_{i=1}^{r} (b_i - a_i)$$

that we will extend into Leb. It is easy to see that μ_0 is well defined and finitely additive. Let λ be the canonical outer measure defined on Σ_0. In our context,

$$\lambda(A) = \inf \left\{ \sum_{i=1}^{r} (b_i - a_i) : \ A \subseteq \bigcup_{i=1}^{r} (a_i, b_i],\ r \geq 1 \right\}$$

where the infimum is on the sets of the form of a disjoint union $\bigcup_{i=1}^{r}(a_i, b_i]$ that contain A. By Theorem 16, the outer measure λ is in fact a measure on $\sigma(\Sigma_0)$. At this point, we could consider that the work is done since a measure has been constructed but it is still not obvious that for $A \in \Sigma_0$, $\mu_0(A) = \lambda(A)$. By finite additivity it is enough to show that $\lambda((a, b]) = b - a$. By construction, we already have that

$$\lambda((a, b]) \leq b - a$$

but for any finite disjoint union of sets $\bigcup_{i=1}^{r}(a_i, b_i]$ such that

$$(a, b] \subseteq \bigcup_{i=1}^{r}(a_i, b_i],$$

we have by simple calculation that

$$b - a \leq \sum_{i=1}^{r}(b_i - a_i)$$

which implies that $b - a \leq \lambda((a, b])$. This reasoning is also applicable to show that $\lambda(\{a\}) = 0$.

15.2 A random variable of given law

The law (or the probability distribution) of a random variable X on the probability triple (Ω, Σ, P) is the image measure $\mathcal{L}_X = P \circ X^{-1}$. For a given $\mathcal{L}(X)$ it is always possible to define a probability triple and a random variable that correspond by taking $X = \text{id}$ and $P = \mathcal{L}_X$. This purely theoretical definition is not that interesting since it does not give any extra information. A more interesting question arises when one imposes a probability triple at the origin (usually $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \text{Leb})$). This new question is tackled by Skorohod construction.

15.2.1 Real valued random variables

Chapter 16

Szemeredi Regularity Lemma

16.1 A basic lemma

A refined version of Cauchy-Schwarz inequality One can use regular Cauchy-Schwarz inequality to obtain the following refined result.

Lemma 41. *Let $(a_i)_{1\leq i\leq n}$ be non-negative, $(b_i)_{1\leq i\leq n} \in \mathbb{R}^n$ and let $b \in \mathbb{R}$ such that*

$$\sum_{i=1}^{n} a_i = 1 \qquad \sum_{i=1}^{n} a_i b_i = b.$$

Let $\mu > 0$ and assume that $\exists j < n$ such that

$$\sum_{i=1}^{j} a_i b_i \geq ab + \mu$$

where $a = \sum_{i=1}^{j} a_i$. Then

$$\sum_{i=1}^{n} a_i b_i^2 \geq b^2 + \frac{\mu^2}{a(1-a)}.$$

Proof. We have that

$$\begin{aligned}
\sum_{i=1}^{n} a_i b_i^2 - b^2 &= \sum_{i=1}^{n} a_i b_i^2 - 2b^2 + b^2 \\
&= \sum_{i=1}^{n} a_i b_i^2 - 2(\sum_{i=1}^{n} a_i b_i b) + \sum_{i=1}^{n} a_i b^2 \\
&= \sum_{i=1}^{j} a_i (b_i - b)^2 + \sum_{i=j+1}^{n} a_i (b_i - b)^2 \\
&\geq \frac{1}{a}\left(\sum_{i=1}^{j} a_i (b_i - b)\right)^2 + \frac{1}{1-a}\left(\sum_{i=j+1}^{n} a_i (b_i - b)\right)^2 \\
&\geq \frac{\mu^2}{a} + \frac{\mu^2}{1-a} = \frac{\mu^2}{a(1-a)}
\end{aligned}$$

where we used that $\sum_{i=1}^{j} a_i(b_i - b) = -\sum_{i=j+1}^{n} a_i(b_i - b)$. □

We can derive a useful corollary:

Corollary 10. *For any sequence $(x_k)_k$ such that*

$$\sum_{k=1}^{m} x_k = \frac{m}{n}\sum_{k=1}^{n} x_k + \delta$$

we have, for $m \leq n$,

$$\sum_{k=1}^{n} x_k^2 \geq \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)^2 + \frac{\delta^2 n}{m(n-m)}.$$

Proof. Use Lemma 41 with $a_i = 1/n$, $\mu = \delta/m$ and $b_i = x_i$. □

16.2 Regular graphs and partitions

In this section, we define the notion of regular graphs that is a graph that has a lot of characteristics in common with a random graph. For a graph $G = (V, E)$ and $X, Y \subset V$, we call **density** between X and Y the quantity

$$d(X, Y) = \frac{e(X, Y)}{|X||Y|}$$

where $e(X, Y)$ is the number of edges between an element of X and an element of Y and $|X|$, $|Y|$ hold for the cardinals of X and Y. Of course, $d(X, Y) \leq 1$ and the equality is obtained if the edges between X and Y correspond to the complete bipartite graph.

Definition 18. *Let $G = (V, E)$ be a graph and let $X, Y \subset V$ be disjoints and non-empty. We say that the pair X, Y is ε-**regular** if for any $A \subseteq X$, $B \subseteq Y$, such that $|A| \geq \varepsilon|X|$ and $|B| \geq \varepsilon|Y|$, it holds that*

$$|d(A, B) - d(X, Y)| \leq \varepsilon.$$

*The pair X, Y is called ε-**irregular** otherwise.*

A **equitable partition** of a graph is defined as $P = (C_0, \ldots, C_k)$ where the number of vertices in $C_1, \ldots, C_k$ are the same. The class C_0 is called the **exceptional class**. The **index** of an equitable partition P is given by

$$\text{Ind } P = \frac{1}{k^2} \sum_{1 \leq i < j \leq k} d(C_i, C_j)^2$$

The index is a suitable notion of the refinement of a partition since we have that $0 \leq \text{Ind } P \leq 1/2$ and $\text{Ind } P \leq \text{Ind } Q$, if Q is a refinement of P.

Definition 19. *Let $G = (V, E)$ be a graph and let P be an equitable partition of V into $C_0, \ldots, C_k$. The partition P is called ε-**regular** if $|C_0| \leq \varepsilon n$ and if at most εk^2 pairs $(C_i, C_j)_{i,j}$ are ε-irregular.*

The important remark in the paper of [11] is that a particular manner to refine irregular partitions ensures that the index increases by a lower bounded quantity and is, then, possible only a finite number of times.

Lemma 42. *Let $G = (V, E)$ be a graph on n vertices and let P be a equitable partition of V into $C_0, \ldots, C_k$. Let ε be such that $4^k > 600\varepsilon^{-5}$. Then, if there is more than εk^2 irregular pairs, there exists a equitable partition Q of size at most $1 + k4^k$ such that the cardinality of the exceptional class does not exceed $|C_0| + \frac{n}{4^k}$ and such that*

$$\textit{Ind } Q \geq \textit{Ind } P + \frac{\varepsilon^5}{20}.$$

We are now able to state the main theorem.

Theorem 18 (Szemeredi Regularity Theorem)**.** *Let $\varepsilon > 0$ and $t \in \mathbb{N}^*$, then there exists integers $N(\varepsilon, t)$ and $M(\varepsilon, t)$ such that every graph $G = (V, E)$ with $|V| \geq N(\varepsilon, t)$, there exists a ε-regular partition of G into $k + 1$ classes such that $t \leq k \leq M(\varepsilon, t)$.*

Proof of Theorem 18. We begin with a trivial partition that have enough elements. Let s be an integer such that $4^s \geq 600\varepsilon^{-5}$, $s \geq t$ and $s \geq 2/\varepsilon$. Define the function f by $f(0) = s$ and for any integer k,

$$f(k+1) = f(k)4^{f(k)}.$$

Let G be a graph (whose number of vertices n is greater than $N(\varepsilon, t)$) and let

$$T = \{k \in \mathbb{N} : \exists \text{ a partition } P \text{ into } 1 + f(k) \text{ classes s.t. Ind } P \geq \frac{k\varepsilon^5}{20} \text{ and } |C_0| \leq \varepsilon n(1 - 2^{-(k+1)})\}.$$

Of course any such partition verify $|C_0| \leq \varepsilon n$ and $0 \in T$ since any partition with $|C_0| \leq \varepsilon n/2$ and letting the rest of C_i being completely free fulfills the assumptions of T. On the other hand, T has a maximum since $\text{Ind } P \leq 1/2$ adn denote k_0 this maximum. Then there exists P a partition into $1 + f(k_0)$ classes such that $\text{Ind } P \geq k_0\varepsilon^5/20$ and $|C_0| \leq \varepsilon n(1 - 2^{-(k_0+1)})$. Assume that P is not a ε-regular partition. Then, by Lemma 42, one can construct another partition P^* into $1 + f(k_0)$ classes such that $\text{Ind } P^* \geq (k_0 + 1)\varepsilon^5/20$. Obvious calculation also show that the exceptional class fulfills the condition of T if

$$\frac{\varepsilon^{-1}}{4^{f(k_0)}} \leq 2^{-(k_0+2)} \Leftarrow 4^s \geq 4\varepsilon^{-1}$$

which is obviously satisfied by the choice of s. This contradict the maximality of k_0 then P is ε-regular. In this construction $M(\varepsilon, t)$ can be taken equal to $f(\lfloor 10\varepsilon^{-5} \rfloor)$ and $N(\varepsilon, t)$ be such that the graph could be cut into $f(M(\varepsilon, t)) + 1$ if needed so $N(\varepsilon, t) = f(M(\varepsilon, t)) + 1$. □

Bibliography

[1] Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Piecewise-polynomial approximations of functions of the classes w_p^α. *Matematicheskii Sbornik*, 115(3):331–355, 1967.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[5] Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic processes and their applications*, 84(2):313–342, 1999.

[6] Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[7] Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[8] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[9] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 2006.

[10] Adrien Saumard. *Estimation par minimum de contraste régulier et heuristique de pente en sélection de modèles*. PhD thesis, 2010. Thèse de doctorat dirigée par Berthet, Philippe Mathématiques et applications Rennes 1 2010.

[11] Endre Szemerédi. Regular partitions of graphs. Technical report, Stanford University, 1975.

[12] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[13] Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[14] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[15] David Williams. *Probability with martingales*. Cambridge university press, 1991.