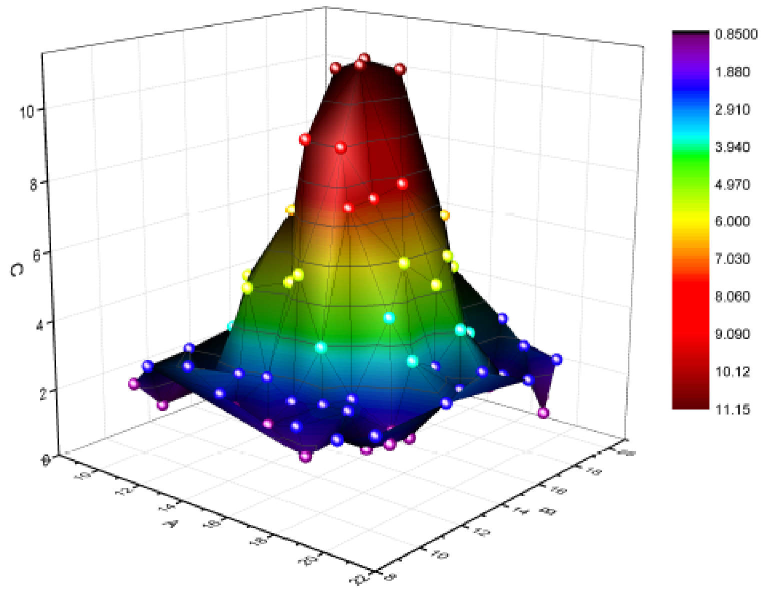
Themes of Statistics

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 $\lim\limits_{T \to +\infty} \overline{\lim} \, \mathbb{P}\left(\|Z_n\| > T k_n\right) = 0$,
- $Z_n = o_P(k_n)$ 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 [18].

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 $\sup_B |\mathbb{P}\left(X_n \in B\right) - \mathbb{P}\left(X \in B\right)| \underset{n\to\infty}{\longrightarrow} 0$, 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 $x, y \in \mathbb{R}^k$
$$cd'(x,y) \le d(x,y) \le Cd'(x,y).$$
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,* $\limsup \mathbb{P}(X_n \in F) \le \mathbb{P}(X \in F)$.

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

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) $\implies$ ii) is obvious since Lipschitz bounded functions are in particular continuous and bounded.
ii) $\implies$ 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,

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

where the last fact holds by monotone convergence.
iii)$\Leftrightarrow$ iv) is obvious by completion.
iii) + iv) $\implies$ 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

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

This chain of inequalities finally imply that

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

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

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

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

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

The previous convergence has to be understood as the sumable (see Definition 18 and Proposition 30) then the sum has only finite number of non zero terms:

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

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

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

We finish by taking the infimum in k.
iii) $\implies$ i) Let $0 < f < 1$ be a continuous function. Using the classical (15.4) and Fatou Lemma, we get

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

We used that the set $\{f(X_n) \ge x\} = \{X_n \in f^{-1}([x, +\infty)\}$ 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 $\mathbb{E}\left[X_n\right] = 1$ 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 $\mathbb{R}^d$.*

$$\begin{tikzcd}
\textit{a.s.} \arrow[rr, bend left, "\textit{Dominated convergence}"] \arrow[dr, Rightarrow] & & \mathbb{L}_p \arrow[dl, Rightarrow] \\
& \mathbb{P} \arrow[ul, bend left, "\exists n_k \textit{ a sub-sequence}"] \arrow[ur, bend right, "\textit{Uniform Integrability}"'] \arrow[d, Rightarrow] & \\
& \textit{(d)} \arrow[u, bend right, "\textit{the limit is constant}"'] &
\end{tikzcd}$$

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 $X_n \xrightarrow{a.s.} X$.

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

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

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

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

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

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 42.
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

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

and

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

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

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

We can use the dominated convergence theorem (see Lemma 32) since $\phi_K(X_n)$ and $\phi_K(X)$ are bounded (and then integrable) to see that there exists n_0 such that $\forall n \geq n_0$,

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

Summing up, we get

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

Then $X_n \xrightarrow{\mathbb{L}_p} X$.

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

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

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

Two exercises about probability convergence

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

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

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* $\forall n \in \mathbb{N}, X_n \leq X_{n+1}$*. 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* $\forall n \geq N$, $\mathbb{E}\left[\|X_n - X\|^p\right] \leq \varepsilon/2^p$*.*

2. *Apply Proposition 31 to show that we can choose* $\delta > 0$ *such that for any* $E \in \mathcal{B}$ *such that* $\mathbb{P}(E) < \delta$*, we have*

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

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

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

(We may use Lemma 26) for $n > N$ *and question 2. for* $n \leq N$*.*

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 $X_1, \ldots, X_n$ 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

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

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 $\forall x \leq y$ and $\forall u \leq v$,

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

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 $x \leq y$ 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

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

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 32

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 17.1.1) 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 $a \leq b$ and $c \leq d$ with $a,b,c,d \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$) corresponding to a event of the form

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

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 $0 \leq F \leq 1$ 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*

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

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

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

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, $F_n(K) \geq \mu_n([-K,K]^k) \geq 1 - \varepsilon$, 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

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

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 $\forall q \in \mathbb{Q}^k$, $H(q) = \lim\limits_{i\to+\infty} F_{n(i,i)}(q)$. 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*

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

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 [11] 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 33). 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 30), $\exists n_0$ 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*

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

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^TZ for any $t\in\mathbb{R}^k$. The random variable t^TZ is normal by definition and $\mathbb{E}\left[t^TZ\right] = t^T\mu$,

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

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 family of $Y_i=(t^TX_i-t^T\mu)_i$ 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. $\sum_{i=1}^{k_n} \operatorname{Cov}(X_{n,i}) \underset{n\to+\infty}{\longrightarrow} \Sigma$

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

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) $\mathbb{E}\left[X_{n,i}\right] = 0$. By the result of Cramer-Wold 1, it is sufficient to show that for all $t \in \mathbb{R}^d$,

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

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,

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

and

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

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,

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

Fix $\underline{\varepsilon_0 > 0}$ and $\varepsilon = \sqrt{\varepsilon_0/2}$. There exists $\underline{N_0}$ such that $\forall n \geq N_0$, $R_n \leq \varepsilon_0/2$. 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, $\forall \delta > 0, \exists n_0, \forall n \geq n_0, \forall i \in \{1, \dots, k_n\}$

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

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 15.1.

Lemma 5. *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*

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

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

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

where we used that for any $z \in \mathbb{C}$ such that $\Re(z) \leq 0$, it holds that $|e^z - 1 - z| \leq |z|^2/2$. See Lemma 27 for a proof of this fact. Using the Taylor-Young formula,

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

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

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

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

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

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 $X_1, \dots, X_n$ 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 $\forall i, j$, $|\mathit{Cov}(X_i, X_j)| \le \phi(|i-j|)$ 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 $\mathrm{Var}(S_n) \to 0$.

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

□

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 $\forall i$, $\mathrm{Var}(X_i) \le c$.

6.1.2 Strong law of large numbers under dependence

One option to prove strong dependent law of large numbers is Lemma 44, 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 [6].

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.

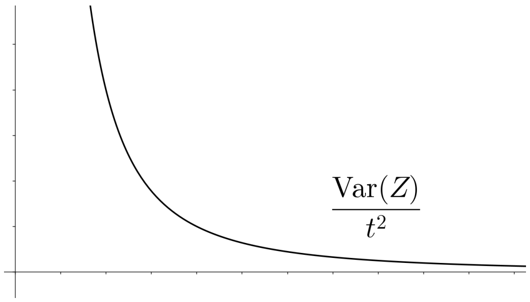


Figure 7.1: The concentration inequality of Bienaymé-Tchebychev

When $Z = Z_n := f_n(X_1, \ldots, X_n)$ is function of independent random variables $X_1, \ldots, X_n$, 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].

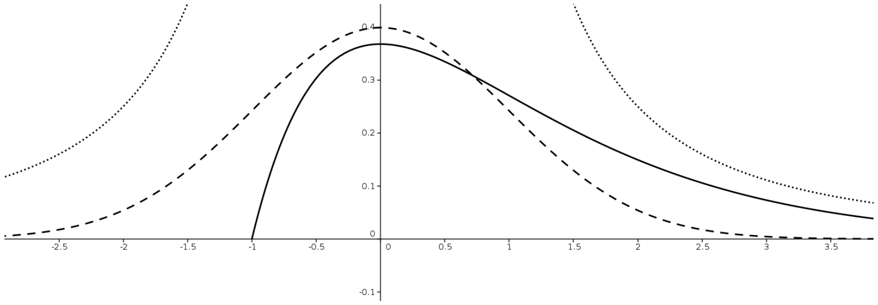


Figure 7.2: In solid line represents the distribution of a variable Z_n. The dotted line is a concentration inequality (here Bienaymé-Tchebychev). The dashed line represents the asymptotic distribution of the variable Z_n.

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 15.

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*

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

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 $\mathcal{G}(\sigma^2)$.*

Definition 5 is only one of some possibilities for the definition of sub-Gaussian random variables. It is actually possible to show that one can characterize those random variables with its moments and also by the existence of a Orlicz norm for the function $\phi(x) = e^{x^2} - 1$ (see Definition 6).

Proposition 14. *Let X be a real valued random variable with $\mathbb{E}\left[X\right] = 0$ then the following are equivalent.*

1. *X is sub-Gaussian.*
2. *There exists $\nu > 0$ such that for all $t > 0$, $\mathbb{P}\left(X \geq t\right) \vee \mathbb{P}\left(-X \geq t\right) \leq e^{-t^2/2\nu}$.*
3. *There exists $C > 0$, s.t. for any integer $q \geq 1$, $\mathbb{E}\left[X^{2q}\right] \leq q!(C)^q$.*
4. *There exists $c > 0$, s.t. $\mathbb{E}\left[e^{X^2/c^2}\right] \leq 2$.*

Exercice 17. *Prove Proposition 14.*

Example 4. *A Gamma random variable Y of parameters (a,b) is sub-Gamma $\Gamma(ab^2,b)$. Indeed, $\mathbb{E}[Y]=ab$ 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, $\log(1-u)-u\le u^2/(2(1-u))$, we have

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

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

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

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

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

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

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

where $g(u)=1+u-\sqrt{1+2u}$ for $u\ge0$. Then

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

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 $0<b\le+\infty$. For any $T\ge0$, 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 $\lambda\in(0,b)$, $\psi(\lambda)\ge\lambda t$. But, then $\psi'(0)\ge t>0$ 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 $t\ge0$,

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

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 16. *Let $Z_1, \dots, Z_N$ 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

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

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

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

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

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

We end by using Lemma 7. □

Using Proposition 16, one can directly derive a bound for the expectation of the maximum of sub-Gamma random variables. If $Z_i \in \Gamma(\nu, c)$,

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

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 $\Psi_Y(\lambda) = \log \mathbb{E}\left[\exp(\lambda Y)\right]$. 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 $X_1, \dots, X_n$ 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*

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

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 17 (Benett inequality)**.** *Let $b > 0$ and $\phi(u) = e^u - u - 1$ for $u \in \mathbb{R}$. Let $X_1, \dots, X_n$ 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 $\mathbb{E}\left[e^{\lambda X_i}\right] - \mathbb{E}[X_i] - 1 \leq \mathbb{E}\left[X_i^2\right]\phi(\lambda)$. Then,

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

□

A even stronger result is the following.

Proposition 18 (Bernstein inequality)**.** *Let $X_1, \dots, X_n$ 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 $S = \sum_{i=1}^n X_i - \mathbb{E}[X_i]$, 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 15. □

7.2 Tensor inequalities and Entropies

The case of sums of independent variables is of course an important case but it does not witnesses the full diversity of the functional of independent variables that one encounters in statistical of probabilistic problems. In this section, we are interested in giving tools that allow to mimic the simple case of sums of independent variables through techniques that we call tensorisation. A tensorization equality of inequality is a result that links the multidimensional case to the one dimensional case. Doing that, one hopes to deduce the concentration of a functional of n independent random variables from the dependency of the functional to each of the variables taken separately. We begin with the simplest tensorization inequality, the so-called Efron-Stein inequality.

7.2.1 Efron-Stein inequality

For a sum of independent random variables $Z = X_1 + \cdots + X_n$, basic calculations give $\mathrm{Var}\left(()\, Z\right) = \sum_i \mathrm{Var}\left(()\, X_i\right)$. In fact, the only needed fact is that the random variables X_i are uncorrelated. The Efron-Stein inequality deals with the case of a generic function of n independent random variables.

Theorem 10 (Efron-Stein)**.** *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X_1, \ldots, X_n)$ for a real valued function f. Assume that $\mathbb{E}\left[Z\right] < \infty$, then*

$$Var(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}^{(i)}(Z)\right]$$

where $\mathrm{Var}^{(i)}(Z) = \mathbb{E}^{(i)}\left[(Z - \mathbb{E}^{(i)}[Z])^2\right]$ *and* $\mathbb{E}^{(i)}[\cdot] = \mathbb{E}\left[\cdot | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\right]$.

Proof. The idea of the proof is to create a martingale $(\Delta_i)_i$ such that Z writes as a sum of the terms Δ_i. We denote by $\mathbb{E}_i\left[\cdot\right]$ the expectation operator conditioned to the variables $X_1, \ldots, X_i$. By convention, we take $\mathbb{E}_0\left[\cdot\right] = \mathbb{E}\left[\cdot\right]$. Then, we define

$$\Delta_i = \mathbb{E}_i\left[Z\right] - \mathbb{E}_{i-1}\left[Z\right],$$

so that we have that

$$Z - \mathbb{E}\left[Z\right] = \sum_{i=1}^{n} \Delta_i.$$

The next step is to show that the random variables Δ_i are uncorrelated so that the variance of Z will equals the sum of the variances of the Δ_i. First of all, $\forall i, j \in \{1, \ldots, n\}$ with $j > i$, $\mathbb{E}_i\left[\Delta_j\right] = \mathbb{E}_i\left[\mathbb{E}_j\left[Z\right]\right] - \mathbb{E}_i\left[\mathbb{E}_{j-1}\left[Z\right]\right] = \mathbb{E}_i\left[Z\right] - \mathbb{E}_i\left[Z\right] = 0$ and then,

$$\mathbb{E}\left[\Delta_i \Delta_j\right] = \mathbb{E}\left[\mathbb{E}_i\left[\Delta_i \Delta_j\right]\right] = \mathbb{E}\left[\Delta_i \mathbb{E}_i\left[\Delta_j\right]\right] = 0.$$

But using Fubini's theorem to obtain that the integration over $X_i, \ldots, X_n$ can be done over X_i first and then over $X_{i+1}, \ldots, X_n$ afterwards, we have that $\mathbb{E}_i\left[\mathbb{E}^{(i)}[Z]\right] = \mathbb{E}_{i-1}\left[Z\right]$. Then,

$$\Delta_i = \mathbb{E}_i\left[Z - \mathbb{E}^{(i)}[Z]\right] \quad \text{and by Jensen's inequality,} \quad \Delta_i^2 \leq \mathbb{E}_i\left[\left(Z - \mathbb{E}^{(i)}[Z]\right)^2\right].$$

So $\mathbb{E}\left[\Delta_i^2\right] \leq \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)}[Z]\right)^2\right] = \mathbb{E}\left[\mathbb{E}^{(i)}\left[\left(Z - \mathbb{E}^{(i)}[Z]\right)^2\right]\right] = \mathbb{E}\left[\mathrm{Var}^{(i)}(Z)\right]$ which gives the result. □

7.2.2 General entropy tensorization

Efron-Stein inequality is actually a special case of a more general fact about a class of functions that define a notion of entropy. The entropies that we define in this section are to be related to Shannon entropy and the related topics. We begin with the result before giving the proper definition of the entropy.

Theorem 11 (Tensorization bound for entropies)**.** *Let $\phi : [0, +\infty) \to \mathbb{R}$ be a continuous and convex function on $[0, +\infty)$. We assume that ϕ is twice differentiable on $(0, +\infty)$ and $\phi'' > 0$ and $1/\phi''$ is concave. Then, the random variable $H_\phi(Z) = \mathbb{E}[\phi(Z)] - \phi(\mathbb{E}[Z])$ satisfy*

$$H_\phi(Z) \leq \mathbb{E}\left[\sum_{i=1}^{n} H_\phi^{(i)}(Z)\right]$$

where $H_\phi^{(i)}(Z) = \mathbb{E}^{(i)}[\phi(Z)] - \phi(\mathbb{E}^{(i)}[Z])$.

Proof. We omit the proof that can be found in [2, Theorem 14.1]. □

The operator H_ϕ is called the **ϕ-entropy** operator. For the specific choice of $\phi = x \log x$, we denote $\mathrm{Ent} = H_\phi$. This last quantity is called the **classical entropy**. It is clear that the function $x \mapsto x \log x$ verifies the conditions of the theorem. It is interesting to note that the choice $x \to x^2$ is also valid and $H_{x \mapsto x^2} = \mathrm{Var}$ is no more than the variance operator. It is an extremal case in a sense since taking $\phi = x^\alpha (\log x)^\beta$ imposes that $1 \leq \alpha \leq 2$.

Exercice 18. *Show that $\phi : x \mapsto x \log x$ fulfills the conditions of Theorem 11.*

7.2.3 Chain rule for various notions of entropy

-Sub-addtitivity of entropy

Theorem 12 (Sub-additivity)**.**

7.2.4 Bounded difference inequalities

7.3 Orlicz norms

In this section, we introduce the notion of Orlicz norm and show its consequences in terms of concentration.

Definition 6. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function such that $\psi(0) = 0$. The **Orlicz norm** of a variable X in $\mathbb{R}^k$ is given by*

$$\|X\|_\psi = \inf\left\{c > 0 : \mathbb{E}\left[\psi\left(\frac{\|X\|}{c}\right)\right] \leq 1\right\}.$$

If the set on the right hand side is empty, we write $\|X\|_\psi = \infty$.

Simple facts over convex functions show that $\psi \geq 0$, is continuous and non-decreasing on $\mathbb{R}_+$. This fact show that the functional $c \mapsto \mathbb{E}[\psi(\|X\|/c)]$ is a continuous non-increasing function so that we can say that the inf in the definition is actually a min. This shows that for $c = \|X\|_\psi$, we have that $\mathbb{E}[\psi(\|X\|/c)] = 1$. The Orlicz norm is not abusively called a norm since we have the following.

Proposition 19. *The operator $\|\cdot\|_\psi$ is a norm over the set of random variables quotiented by the relation $\mathcal{R}$ given by $X \underset{\mathcal{R}}{\sim} Y$ if and only if $X = Y$ almost surely.*

Proof. Let X and Y be two random vectors with finite Orlicz norms. Let $c_1 > \|X\|_\psi$ and $c_2 > \|Y\|_\psi$.

$$\begin{aligned}
\mathbb{E}\left[\psi\left(\frac{\|X+Y\|}{c_1+c_2}\right)\right] &\leq \mathbb{E}\left[\psi\left(\frac{\|X\|}{c_1}\frac{c_1}{c_1+c_2} + \frac{\|Y\|}{c_2}\frac{c_2}{c_1+c_2}\right)\right] \\
&\leq \frac{c_1}{c_1+c_2}\mathbb{E}\left[\psi\left(\frac{\|X\|}{c_1}\right)\right] + \frac{c_2}{c_1+c_2}\mathbb{E}\left[\psi\left(\frac{\|Y\|}{c_2}\right)\right] \\
&= \frac{c_1}{c_1+c_2} + \frac{c_2}{c_1+c_2} = 1.
\end{aligned}$$

Since this is true for any $c_1 > \|X\|_\psi$ and $c_2 > \|Y\|_\psi$, this shows that $\|X+Y\|_\psi \leq \|X\|_\psi + \|Y\|_\psi$. From the same kind of calculations, one can show that $\|\lambda X\|\psi = |\lambda|\|X\|_\psi$. Now assume that $\|X\|_\psi = 0$ and assume that $X = 0$ a.s. is false. Then one can find $\delta > 0$ such that $p = \mathbb{P}(\|X\| \geq \delta) > 0$. Since $\|X\|_\psi = 0$, we have that for any $c > 0$ that $\mathbb{E}[\psi(\|X\|/c)] \leq 1$. Then

$$\mathbb{E}\left[\psi\left(\frac{\|X\|}{c}\right)\right] = \int_0^\infty \mathbb{P}\left(\|X\| \geq c\psi^{-1}(t)\right) dt \leq 1.$$

But since ψ^{-1} is non-decreasing, when one takes $c\psi^{-1}(t) \leq \delta$, we see that $\mathbb{P}\left(\|X\| \geq c\psi^{-1}(t)\right) \geq p$ which implies that the above integral is infinite, absurd. Then $X = 0$ a.s. and so $X \underset{\mathcal{R}}{\sim} 0$ which concludes the proof. □

Orlicz norms ψ_p The important examples are the power functions and the exponential type functions. If $\psi(x) = x^p$, then $\|X\|_\psi = \|X\|_p$. When $\phi(x) = e^{x^2} - 1$, for any $c \geq \|X\|_\psi$, we have that

$$\mathbb{E}\left[e^{\frac{X^2}{c^2}}\right] \leq 2 \tag{7.3}$$

and then by Proposition 14, the random variable X/c is sub-Gaussian. The value $\|X\|_\psi$ is then the smallest c such that (7.3) holds. Finally, we denote by $\psi_p(x) = e^{x^p} - 1$ and by $\|\cdot\|_{\psi_p}$ the corresponding Orlicz norm. It is immediate to see that for any random variable X,

$$\|X\|_p \leq \|X\|_{\psi_p} \tag{7.4}$$

since we have the inequalities $\psi_p(x) \geq x^p$ for any $p \geq 1$.
Where the Orlicz norms for $\psi(x) = x^p$ control the existence of moments of order p, the Orlicz norms ψ_p control the exponential concentration of the random variables. The precise statement is as follows.

Proposition 20. *Let X be a random variable and let $p \in [1, +\infty)$. The following facts are equivalents*

1. $\|X\|_{\psi_p} < \infty$.
2. *There exist $C, K > 0$ such that*

$$\mathbb{P}(\|X\| > t) \leq Ke^{-Ct^p} \qquad \forall t > 0.$$

If 1. occurs then 2. is verified with $C = \|X\|_{\psi_p}^{-p}$ and $K = 2$. If 2. occurs then $\|X\|_{\psi_p} \leq ((1+K)/C)^{1/p}$.

Proof. Assume that *1.* holds. Then

$$\begin{aligned}\mathbb{P}(\|X\| > t) &\leq \mathbb{P}\left(\psi_p\left(\frac{\|X\|}{\|X\|_{\psi_p}}\right) \geq \psi_p\left(\frac{t}{\|X\|_{\psi_p}}\right)\right)\\ &\leq 1 \wedge \frac{1}{\psi_p\left(\frac{t}{\|X\|_{\psi_p}}\right)}\\ &\leq 2\exp\left(-\frac{t^p}{\|X\|_{\psi_p}^p}\right),\end{aligned}$$

where we used that $\forall u > 0$, it holds that $1 \wedge (e^u - 1)^{-1} \leq 2e^{-u}$. Now it *2.* holds,

$$\begin{aligned}\mathbb{E}\left[e^{\|X\|^p/c^p} - 1\right] &= \mathbb{E}\left[\int_0^{\|X\|^p} \frac{e^{s/c^p}}{c^p} ds\right]\\ &= \int_0^{+\infty} \mathbb{P}\left(\|X\| > s^{1/p}\right) \frac{e^{s/c^p}}{c^p} ds\\ &\leq \int_0^{+\infty} Ke^{-Cs} \frac{e^{s/c^p}}{c^p} ds\\ &= \frac{K}{c^p} \frac{1}{C - 1/c^p}.\end{aligned}$$

Then, when $c \geq ((1+K)/C)^{1/p}$,

$$\frac{K}{c^pC - 1} \leq \frac{K}{((1+K)/C)C - 1} = 1,$$

and so $\|X\|_{\psi_p} \leq c$ which gives the result. □

Proposition 21. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function such that $\psi(0) = 0$ and there exists $c > 0$ such that*

$$\limsup_{x,y\to\infty} \frac{\psi(x)\psi(y)}{\psi(cxy)}.$$

Then, for $X_1, \dots, X_N$ real random variables, it holds that

$$\| \max_{1\le i\le N} |X_i| \|_\psi \le K\psi^{-1}(N) \max_{1\le i\le N} \|X_i\|_\psi.$$

where $K = K(\psi)$ is a constant only depending on ψ.

Proof. We first assume that $\psi(1) \le 1/2$ and that for all $x, y \ge 1$, $\psi(x)\psi(y) \le \psi(cxy)$. In particular for any $x \ge y \ge 1$, $\psi(x/y) \le \psi(cx)/\psi(y)$. Now, let $y \ge 1$ and $u > 0$ so it holds that

$$\begin{aligned}\max_{1\le i\le N} \psi\left(\frac{|X_i|}{uy}\right) &\le \max_{1\le i\le N} \left[\frac{\psi(c|X_i|/u)}{\psi(y)} \mathbb{1}_{\frac{|X_i|}{uy}\ge 1} + \psi\left(\frac{|X_i|}{uy}\right) \mathbb{1}_{\frac{|X_i|}{uy}<1}\right] \\ &\le \sum_{i=1}^N \frac{\psi(c|X_i|/u)}{\psi(y)} + \psi(1).\end{aligned}$$

So taking, $u = c\max_{1\le i\le N} \|X_i\|_\psi$ and $y = \psi^{-1}(2N)$, one obtains

$$\mathbb{E}\left[\psi\left(\frac{\max_{1\le i\le N}|X_i|}{uy}\right)\right] \le \frac{N}{\psi(y)} + \frac{1}{2} \le 1.$$

This shows that $\| \max_{1\le i\le N}|X_i| \|_\psi \le c\psi^{-1}(2N) \max_{1\le i\le N} \|X_i\|_\psi$, but since ψ is convex, the generalized inverse ψ^{-1} is concave and $\psi^{-1}(0) = 0$, then $\psi^{-1}(2N) \le 2\psi^{-1}(N)$ and we get the result in this special case with $K = 2c$. Going back to the general case, under the conditions of the theorem on ψ, one can always find $0 < \sigma \le 1$ and $\tau > 0$ such that $\phi(x) = \sigma\psi(\tau x)$,

$$\forall x, y \ge 1,\ \phi(x)\phi(y) \le \phi(cxy) \qquad \text{and} \qquad \phi(1) \le \frac{1}{2}.$$

Then the same concavity fact gives that $\forall u > 0$, $\phi^{-1}(u) \le \psi^{-1}(u)/\sigma\tau$ which is equivalent to saying that $\forall u > 0$, $\psi(u\sigma\tau) \le \phi(u)$. Then for any $c > \|X\|_\phi$, we have that

$$\mathbb{E}\left[\psi\left(\frac{|X|\sigma\tau}{c}\right)\right] \le \mathbb{E}\left[\phi\left(\frac{|X|}{c}\right)\right] \le 1$$

which imply that $\|X\|_\psi \le \|X\|_\phi/\sigma\tau$. But, for all $u > 0$,

$$\mathbb{E}\left[\phi\left(\frac{|X|}{\tau u}\right)\right] = \mathbb{E}\left[\sigma\psi\left(\frac{|X|}{u}\right)\right] \le \mathbb{E}\left[\psi\left(\frac{|X|}{u}\right)\right],$$

which shows that $\|X\|_\phi/\tau \le \|X\|_\psi$. Using these two inequalities, we have that

$$\Big\| \max_{1\le i\le N} |X_i| \Big\|_\psi \le \frac{1}{\sigma\tau} \Big\| \max_{1\le i\le N} |X_i| \Big\|_\phi \le \frac{2c}{\sigma\tau}\phi^{-1}(N) \max_{1\le i\le N} \|X_i\|_\phi \le \frac{2c\tau}{(\sigma\tau)^2}\psi^{-1}(N) \max_{1\le i\le N} \|X_i\|_\psi.$$

Then Proposition 21 holds with $K = \frac{2c}{\sigma^2\tau}$ which depends only on ψ. □

Exercice 19. *Show that for any convex function ψ satisfying the hypothesis of Proposition 21, one can find ϕ, σ and τ such that $\phi(x) = \sigma\psi(\tau x)$,*

$$\forall x, y \ge 1,\ \phi(x)\phi(y) \le \phi(cxy) \qquad \textit{and} \qquad \phi(1) \le \frac{1}{2}.$$

Exercice 20. *Show that ψ_p satisfy the hypothesis of Proposition 21 with $c = 1$, $\tau = 1$ y $\sigma = 1/(2(e-1))$. Deduce that in that case, Proposition 21 holds with $K = 8(e-1)^2 \le 24$.*

The sub-Gaussian case (ψ_2) When one choose $\psi = \psi_2$ and if the random variables $X_1, \dots, X_N$ are sub-Gaussian $\mathcal{G}(\nu)$, the result gives is that

$$\| \max_{1\le i\le N} |X_i| \|_{\psi_2} \le K\sqrt{\log(N+1)} \max_{1\le i\le N} \|X_i\|_{\psi_2} \le K\sqrt{2\nu\log(N+1)}.$$

The bound on the right hand side is exactly of the same form as the result of Proposition 16. As a consequence, when the random variables X_i are sub-Gaussian, both the expected value of the maximum and the Orlicz norm are controled by a bound propostional to $\sqrt{\nu\log N}$. By a trivial manipulations, one finally have that there exist positive constants K and C

$$\mathbb{P}\left(\max_{1\le i\le N} |X_i| > t\right) \le Ke^{-C\frac{t^2}{\nu\log N}}. \tag{7.5}$$

Sub-optimality As good as (7.5) seems, the bound is actually not optimal. The Orlicz norm does not takes into account the possible bias that the maximum introduces. Indeed, if $X_1, \dots, X_N$ are sub-Gaussian random variables of constant ν, one can show that

$$\mathbb{P}\left(\max_{1\leq i\leq N} |X_i| > t\right) \leq 2Ne^{-\frac{t^2}{2\nu}} = 2e^{-\frac{t^2}{2\nu}+\log(N)}. \tag{7.6}$$

The term $\log(N)$ is actually more a bias term in this last inequality than a multiplicative term as in (7.5). The bound (7.5) is the best bound that one can obtain with no bias term inside the exponential. In the sequel we will use (7.5) or (7.6) depending on the purpose one is willing to achieve. It is however wrong to think that the bias and the tail of the maximum are of the same order in the case of Gaussian random variables. Indeed the tail behavior is actually not depending on the number N in the maximum and is only dependent on the supremum of the variances.

Proposition 22. *Let $X = (X_1, \dots, X_N)$ be centered Gaussian vector with $\sigma^2 \geq \max \mathbb{E}\left[X_i^2\right]$, then for any $t > 0$ we have*

$$\mathbb{P}\left(\left|\max_{1\leq i\leq N} X_i - \mathbb{E}\left[\max_{1\leq i\leq N} X_i\right]\right| \geq t\right) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right)$$

Proof. Let Γ be the covariance matrix of the Gaussian vector $X = (X_1, \dots, X_N)$ and let $Y = (Y_1, \dots, Y_N)$ be a Gaussian vector of i.i.d. standard Gaussian variables. Since Γ is a positive semidefinite matrix, one can define its square root matrix $\sqrt{\Gamma}$. By construction, the vector X and the vector $\sqrt{\Gamma}Y$ have the same distribution. Let $f(Y) = \max_{i=1,\dots,N}(\sqrt{\Gamma}Y)_i$. It remains to prove the concentration bound on the function f. But for any two vectors $u, v \in \mathbb{R}^N$, and $i \in \{1, \dots, N\}$ we have

$$|(\sqrt{\Gamma}u)_i - (\sqrt{\Gamma}v)_i| = \left|\sum_{j=1}^N (\sqrt{\Gamma})_{i,j}(u_j - v_j)\right| \leq \left(\sum_{j=1}^N (\sqrt{\Gamma})_{i,j}^2\right)^{1/2} \|u-v\|_2 = \Gamma_{i,i}^{1/2}\|u-v\|_2 = \sqrt{\mathrm{Var}\left(X_i\right)}\|u-v\|_2,$$

but then

$$|f(u) - f(v)| \leq \max_{1\leq i\leq N} |(\sqrt{\Gamma}u)_i - (\sqrt{\Gamma}v)_i| \leq \sigma\|u-v\|_2.$$

Therefore, the function f is Lipschitz of constant σ and Theorem 16 applies. □

7.3.1 Gaussian concentration inequality

This section shows one important result over Lipschitz functions of independent Gaussian variables. The main theorem rely on the approximation of a Gaussian random variable by a sum of Rademacher random variables. This approximation will allow us to prove a logarithmic Sobolev inequality inherited from the specific behavior of functions on the binary hypercube. This subject is vast and has given a lot of interesting consequences for concentration and isoperimetric problems. The set $\{-1,1\}^n$ is called the **binary hypercube** of dimension n. A **Rademacher** random variable is a random variable X on $\{-1,1\}$ such that $\mathbb{P}(X=-1) = \mathbb{P}(X=1) = 1/2$. For a function f on the hypercube, the **discrete derivative** in the i-th coordinate is given by

$$\nabla_i f(x) = \frac{f(x) - f(\overline{x}^{(i)})}{2}$$

where $\overline{x}^{(i)} = (x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)$ and the **discrete gradient** is the vector $\nabla f(x) = (\nabla_1 f(x), \dots, \nabla_n f(x))$. A logarithmic Sobolev inequality is a result that bounds the entropy of a random vector with its variance. More precisely, we have the following theorem.

Theorem 13 (Logarithmic Sobolev on the hypercube)**.** *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $X \in \{-1,1\}^n$ be a vector of i.i.d. Rademacher random variables. Then,*

$$\mathrm{Ent}(f^2(X)) \leq 2\mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

Note that if one applies the Efron-Stein inequality in this context, one obtains $\mathrm{Var}\left(f(X)\right) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$. But if f is a non-negative function, $\mathrm{Var}\left(f(X)\right) \leq \mathrm{Ent}(f^2(X))$ which shows that Theorem 13 is stronger than Theorem 10.

Exercice 21. *Let $\phi_p(Z) = (\mathbb{E}\left[Z^2\right] - (\mathbb{E}\left[Z^p\right])^{2/p})/(1/p - 1/2)$ for $p \in [1,2)$. Show that for a non-negative random variable Z, the function $p \mapsto \phi_p(Z)$ is non-decreasing. Calculate $\phi_1(Z)$ and $\lim_{p\to 2}\phi_p(Z)$ and deduce that $Var(Z) \leq \mathrm{Ent}(Z^2)$.*

Proof of Theorem 13. Theorem 12 for the random variable $f^2(X)$ gives

$$\mathrm{Ent}(f^2(X)) \leq \mathbb{E}\left[\sum_{i=1}^n \mathrm{Ent}^{(i)}(f^2(X))\right]$$

where $\mathrm{Ent}^{(i)}$ holds for the entropy with respect to the measure conditioned on $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. In the following, we show that

$$\mathrm{Ent}^{(i)}(f^2(X)) \leq 2\mathbb{E}^{(i)}[(\nabla_i f(X))^2]$$

which is a one dimensional problem. Then it is enough to show that for a function $f : \{-1, 1\} \to \mathbb{R}$,

$$\mathrm{Ent}(f^2(X)) \leq \frac{1}{2}\mathbb{E}\left[(f(1) - f(-1))^2\right].$$

Then denoting $a = f(1)$ and $b = f(-1)$, the last inequality rewrites

$$\frac{a^2}{2}\log(a^2) + \frac{b^2}{2}\log(b^2) - \frac{a^2+b^2}{2}\log\left(\frac{a^2+b^2}{2}\right) \leq \frac{(a-b)^2}{2}.$$

We are now reduced to prove an elementary inequality for all $a, b \in \mathbb{R}$. First of all, we can assume that $0 \leq b \leq a$ by symmetry and since $(|a| - |b|)^2 \leq (a-b)^2$. Now fix $b \geq 0$ and let

$$h(x) = \frac{x^2}{2}\log(x^2) + \frac{b^2}{2}\log(b^2) - \frac{x^2+b^2}{2}\log\left(\frac{x^2+b^2}{2}\right) - \frac{(x-b)^2}{2}$$

from which we see that $h'(b) = 0$ and $h''(x) \leq 0$. The function h is then concave with derivative 0 in b, then h is non-positive on the interval $[b, +\infty)$ which proves the inequality. □

Lemma 9. *Assume that for a random variable Z, there exists a constant $\nu > 0$ such that we have $\forall \lambda > 0$,*

$$\mathrm{Ent}(e^{\lambda Z}) \leq \frac{\lambda^2 \nu}{2}\mathbb{E}\left[e^{\lambda Z}\right]$$

then Z is sub-Gaussian $\mathcal{G}(\nu)$ and so, for example, $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2e^{-t^2/2\nu}$, for all $t \geq 0$.

Proof. Let $F(\lambda) = \mathbb{E}\left[e^{\lambda Z}\right]$ then $F'(\lambda) = \mathbb{E}\left[Ze^{\lambda Z}\right]$. The condition of the theorem rewrites

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \frac{\lambda^2\nu}{2}F(\lambda)$$

or again

$$\left(\frac{\log F(\lambda)}{\lambda}\right)' \leq \frac{\nu}{2}.$$

By elementary integration on the interval $(0, \lambda]$,

$$\frac{\log F(\lambda)}{\lambda} - \frac{F'(0)}{F(0)} = \frac{\log F(\lambda)}{\lambda} - \mathbb{E}[Z] \leq \frac{\nu\lambda}{2}$$

which finally gives $\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \lambda^2\nu/2$ and then $Z - \mathbb{E}[Z]$ is sub-Gaussian of constant ν. □

A direct consequence of the previous result is the following concentration inequality.

Theorem 14. *Let $X = (X_1, \dots, X_n)$ be i.i.d. random variables of Rademacher and let $f : \{-1, 1\} \to \mathbb{R}$ be a real valued function such that there exists a constant $\sigma > 0$ such that $\|\nabla f(x)\| \leq \sigma$ for all $x \in \{-1, 1\}^n$. Then $f(X)$ is a sub-Gaussian random variable of constant $2\sigma^2$ so that for all $t \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2e^{-\frac{t^2}{4\sigma^2}}.$$

Proof. Theorem 13 with the function $g = e^{\lambda f/2}$ gives $\mathrm{Ent}(e^{\lambda f(X)}) \leq 2\mathbb{E}\left[(\nabla e^{\lambda f(X)/2})^2\right]$. But since for any $a, b \in \mathbb{R}$ such that $a \geq b$,

$$(e^a - e^b) \leq \frac{(a-b)}{2}e^a,$$

we have that

$$\mathbb{E}\left[(\nabla_i e^{\lambda f(X)/2})^2\right] = 2\mathbb{E}\left[(\nabla_i e^{\lambda f(X)/2})_+^2\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[(\nabla_i f(X))_+^2 e^{\lambda f(X)}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[(\nabla_i f(X))^2 e^{\lambda f(X)}\right]$$

which implies that $\mathrm{Ent}(e^{\lambda f(X)}) \leq \lambda^2\sigma^2\mathbb{E}\left[e^{\lambda f(X)}\right]$. We finish by using Lemma 9 for $\nu = 2\sigma^2$. □

Theorem 13 can be used to prove the same kind of result for vectors of independent Gaussian random variables. This is the subject of the next theorem.

Theorem 15. *Let X be a centered Gaussian vector in $\mathbb{R}^n$ of covariance matrix I_n and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of $\mathcal{C}_1(\mathbb{R}^n)$. Then*

$$\operatorname{Ent}(f^2(X)) \leq 2\mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

Proof. By the sub-additivity of entropy of Theorem 12, it is enough to prove the theorem in dimension 1 and show that

$$\operatorname{Ent}^{(i)}(f^2(X)) \leq 2\mathbb{E}^{(i)}[\partial_i f(X)^2].$$

So, let $f : \mathbb{R} \to \mathbb{R}$ continuously differentiable. The idea is to use that if $X \sim \mathcal{N}(0,1)$ then for a sequence $(\varepsilon_i)_i$ of Rademacher random variables,

$$X_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xrightarrow{(d)} X$$

So by the continuous transformation property (of Theorem 1) we have that

$$\operatorname{Ent}\left(f^2\left(X_n\right)\right) \underset{n\to+\infty}{\longrightarrow} \operatorname{Ent}(f^2(X)).$$

But

$$\operatorname{Ent}\left(f^2\left(X_n\right)\right) \leq 2\mathbb{E}\left[\sum_{j=1}^n \left(\frac{f\left(X_n\right) - f\left(X_n - \frac{2\varepsilon_j}{\sqrt{n}}\right)}{2}\right)^2\right] = 2\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \left(\frac{f\left(X_n\right) - f\left(X_n - \frac{2\varepsilon_j}{\sqrt{n}}\right)}{2\varepsilon_j/\sqrt{n}}\right)^2\right] \underset{n\to+\infty}{\longrightarrow} 2\mathbb{E}\left[f'(X)^2\right]$$

since f is $\mathcal{C}_1(\mathbb{R})$. □

The main consequence of Theorem 15 is a concentration theorem for Lipschitz functions of independent random variables.

Theorem 16 (Tsirelson-Ibragimov-Sudakov)**.** *Let $(X_1, \dots, X_n)$ be independent Gaussian random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function of constant L. Then the random variable $Z = f(X_1, \dots, X_n)$ is sub-Gaussian of constant L^2 that is $\forall \lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq e^{\frac{\lambda^2 L^2}{2}}.$$

In particular, for any $t > 0$,

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \geq t\right) \leq e^{-\frac{t^2}{2L^2}}.$$

Proof. We first assume that f is continuously differentiable and $\|\nabla f\|_\infty \leq L$. We can assume without loss of generality that Z is centered $\mathbb{E}\left[Z\right] = 0$. We apply Theorem 15 with the function $e^{\lambda f/2}$, then

$$\operatorname{Ent}(e^{\lambda f}) \leq 2\mathbb{E}\left[\|\nabla e^{\lambda f/2}\|^2\right] = \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda f}\|\nabla f\|^2\right] \leq \frac{\lambda^2 L^2}{2}\mathbb{E}\left[e^{\lambda f}\right].$$

Finally using Lemma 9 we get the result. □

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 $X_1, \ldots, X_n$. 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

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

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 (17.2). 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 of empirical processes

8.2.1 Education vs Employment

In our model a population of individuals $X_1 = (Y_1, Z_1), \ldots, X_n = (Y_n, Z_n)$ 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 23. *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 7. *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 7, 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 8. *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 $i \in \{1,\ldots,N\}$ 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 22. *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 24. *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 $(c_j)_{j=1,\ldots,N}$ of elements of A such that

- For any $i \neq j$, $\|c_i - c_j\| \geq \varepsilon$,
- The set $\{c_1,\ldots,c_N\}$ 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 25. *A ε-net $(c_i)_{i=1,\ldots,N}$ of a set A forms the centers set Centers$(\mathcal{C})$ of a covering $\mathcal{C}$ of A.*

Proof. Let $(c_i)_{j=1,\ldots,N}$ 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 $\{c_1,\ldots,c_N\} \cup \{x\}$ is also an ε-net of A which contradicts the maximality of the initial ε-net $\{c_1,\ldots,c_N\}$. □

Lemma 10. *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 $(c_i)_{j=1,\ldots,N}$ be a ε-net of the ball $B_d(R)$. By Proposition 25, 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 10. 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 11)*

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

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

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 17. *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

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

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 11. *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 24. 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 18), this necessary condition will be assumed.

Proposition 26. *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 39, 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 18. *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 17 in two perspectives. First, the condition holds on a notion of entropy that is smaller since, by Proposition 24 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 17) 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 18. 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 12. *Let $X_1, \dots, X_n$ 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 $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher random variables and independent from $X_1, \dots, X_n$, 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 12 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 13. *Assume that for any function $f \in \mathcal{F}$ and a $\delta > 0$,*

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

Then, it holds that

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

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 $A_f = \{\mathbf{X} : \ |\int fd(P_n - P)| > \delta\}$. 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

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

□

To get a result of the form of Lemma 12, one can apply the Rademacher random variables trick and get the following result.

Corollary 5. *Let $\varepsilon_1, \dots, \varepsilon_n$ be Rademacher random variables independent from the X_i. Then, under the hypothesis of Lemma 13,*

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

Proof. The proof used the same ideas as the proof of Lemma 12 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 [8]. 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 [7] and further studied and extended by Talagrand (see [14]).

Lemma 14. *Let $X_1, \dots, X_N$ 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*

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

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

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

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

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

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 19 (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 14 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 19 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 entire set $\mathcal{T}$. 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 19, we get that one can construct a continuous version of the process $(X_t)_t$ when the entropy integral is finite. In fact, Theorem 19 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

For this section, we follow the excellent [9]. The subject of this part is to understand the specific case of gaussian processes for which we show that entropic lower bounds are achievable. This complete argument is called Sudakov minoration. This part uses intensively the ideas behind the so-called comparison theorems where Slepian's Lemmas form the key stone of this section. We begin with a simple lower bound of the maximum of a collection of independent Gaussian random variables.

Proposition 27. *Let $N \geq 6$. Let $X_1, \dots, X_N$ be i.i.d. standard Gaussian random variables. Then, there exists a universal constant $C > 0$ such that*

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

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. Since the variables $X_1, \dots, X_n$ are centered, we see that $\mathbb{E}\left[\max_{i=1,\dots,N} X_i\right] = \mathbb{E}\left[\max_{i=1,\dots,N}(X_i - X_1)\right]$ where the last inequality show that the the quantity is non-negative since $\max_i X_i \geq X_1$. For any $\delta > 0$, one has that

$$\mathbb{E}\left[\max_{i=1,\dots,N} X_i - X_1\right] = \int_0^\infty \mathbb{P}\left(\max_{i=1,\dots,N} X_i - X_1 > t\right) dt \geq \int_0^\delta \mathbb{P}\left(\max_{i=1,\dots,N} X_i - X_1 > \delta\right) dt = \delta[1-(1-\mathbb{P}\left(X_2 - X_1 > \delta\right))^{N-1}].$$

so that if we choose δ such that $\mathbb{P}\left(X_2 - X_1 > \delta\right) \geq 1/(N-1)$, we have that

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

But $\mathbb{P}\left(X_2 - X_1 > \delta\right) = \mathbb{P}\left(\mathcal{N} > \frac{\delta}{\sqrt{2}}\right) = 1/2 - \mathbb{P}\left(0 < \mathcal{N} < \frac{\delta}{\sqrt{2}}\right)$ where $\mathcal{N}$ is a standard Gaussian variable. But

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

so we have that $\mathbb{P}\left(X_2 - X_1 > \delta\right) \geq 1/2(1 - \sqrt{1-e^{-\delta^2/2}})$. The condition on δ is verified if one takes

$$\frac{1}{2}(1 - \sqrt{1-e^{-\delta^2/2}}) \geq \frac{1}{N-1}$$

which is verified for $\delta = \sqrt{2\log((N-1)/4)}$. We finally have that

$$\mathbb{E}\left[\max_{i=1,\dots,N} X_i\right] \geq (1-e^{-1})\sqrt{2\log((N-1)/4)}.$$

Finally, one can conclude the first fact of the Proposition by showing that for there exists a constant C such that $(1-e^{-1})\sqrt{2\log((N-1)/4)} \geq C\sqrt{\log N}$. The constant e^{-1} can be reduced by taking for example $\delta = \sqrt{2\log(p(N-1)/4)}$ the bound becomes $(1-e^{-p})\sqrt{2\log(p(N-1)/4)}$ and we get

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

by taking $p \to \infty$. □

In the proof of Theorem 21 we will use the result of Proposition 27 in the form

$$\mathbb{E}\left[\max_{i=1,\dots,N} X_i\right] \geq \frac{1}{\sqrt{2}}\sqrt{\log(N)} \tag{8.6}$$

where this sub-optimal result can be optained with simpler calculations than in Proposition 27.

Comparison Theorems

The comparison theorems deal with the domination of the probabilities of some events of a gaussian vector X by the same probability for another gaussian vector Y such that Y dominates X in a certain sense. The meaning that one impose behind the domination can take various forms. In the sequel, we mainly deal with the case of domination in the covariance structure.

Theorem 20. *Let $X = (X_1, \dots, X_N)$ and $Y = (Y_1, \dots, Y_N)$ be two centered Gaussian vectors in $\mathbb{R}^N$ such that*

$$\begin{aligned}
\mathbb{E}\left[X_iX_j\right] &\leq \mathbb{E}\left[Y_iY_j\right] \quad \textit{for } (i,j) \in A \\
\mathbb{E}\left[X_iX_j\right] &\geq \mathbb{E}\left[Y_iY_j\right] \quad \textit{for } (i,j) \in B \\
\mathbb{E}\left[X_iX_j\right] &= \mathbb{E}\left[Y_iY_j\right] \quad \textit{for } (i,j) \notin A \cup B
\end{aligned}$$

for A and B being disjoints subsets of $\{1,\dots,N\}^2$. Let f be a function on $\mathbb{R}^N$ such that its second derivatives in the weak sense of distributions are such that

$$\begin{aligned}\partial_{i,j} f &\geq 0 \quad \textit{for } (i,j)\in A\\ \partial_{i,j} f &\leq 0 \quad \textit{for } (i,j)\in B.\end{aligned}$$

Then

$$\mathbb{E}\left[f(X)\right] \leq \mathbb{E}\left[f(Y)\right].$$

Proof. Since the conclusion of the theorem is purely in expectation, one can assume that X and Y are independent and consider for $t\in[0,1]$ the random variable $Z(t) = (1-t)^{1/2}X + t^{1/2}Y$. Denote by $\phi(t) = \mathbb{E}\left[f(Z(t))\right]$ so that one can differentiate ϕ to get

$$\phi'(t) = \sum_{i=1}^N \mathbb{E}\left[\partial_i f(Z(t)) Z_i'(t)\right],$$

where $Z_i'(t) = Y_i/(2\sqrt{t}) - X_i/(2\sqrt{1-t})$. Using Stein identity in Proposition 36 for the function $F = \partial_i f$ and on the Gaussian variable X, Y we get that

$$\mathbb{E}\left[\partial_i f(Z(t))Z_i'(t)\right] = \sum_{j=1}^N \left(\frac{\sqrt{t}}{2\sqrt{t}}\mathbb{E}\left[Y_iY_j\right] - \frac{\sqrt{1-t}}{2\sqrt{1-t}}\mathbb{E}\left[X_iX_j\right]\right)\mathbb{E}\left[\partial_{i,j}f(Z(t))\right] = \frac{1}{2}\sum_{j=1}^N(\mathbb{E}\left[Y_iY_j\right] - \mathbb{E}\left[X_iX_j\right])\mathbb{E}\left[\partial_{i,j}f(Z(t))\right].$$

Then in the matricial notations, we get that

$$\phi'(t) = \frac{1}{2}\operatorname{Tr}\left(\mathbb{E}\left[\nabla^2 f(Z(t))\right](\Sigma_Y - \Sigma_X)\right). \tag{8.7}$$

The condition of the theorem imply that $\phi'(t)\geq 0$ and then $\mathbb{E}\left[f(X)\right] = \phi(0)\leq\phi(1) = \mathbb{E}\left[f(Y)\right]$. □

Theorem 20 has important consequences as Slepian lemma that allows to upper bound the maximum of a collection of Gaussian random variables with the maximum over another collection of Gaussian variables of greater covariances. It is also possible to extract more information of Equation (8.7) to quantify the difference of the values of the maxima when one controls the difference in the covariance matrices.

Lemma 15 (Slepian 1). *Let $X = (X_1,\dots,X_N)$ and $Y = (Y_1,\dots,Y_N)$ be two centered Gaussian vectors in $\mathbb{R}^N$ such that*

$$\begin{aligned}\mathbb{E}\left[X_iX_j\right] &\leq \mathbb{E}\left[Y_iY_j\right] \quad \textit{for all } i\neq j\\ \mathbb{E}\left[X_i^2\right] &= \mathbb{E}\left[Y_i^2\right] \quad \textit{for all } i.\end{aligned}$$

Then for all real numbers $\lambda_1,\dots,\lambda_N$,

$$\mathbb{P}\left(\bigcup_{i=1}^N\{Y_i>\lambda_i\}\right) \leq \mathbb{P}\left(\bigcup_{i=1}^N\{X_i>\lambda_i\}\right).$$

This lemma is the main ingredient to show Markovian-like results on stationary Gaussian processes. For example, one has the following consequence.

Exercice 23. *Let $(X_t)_{t\geq 0}$ is a stationary Gaussian process such that for any $t\geq 0$, $\mathbb{E}\left[X_tX_0\right]\geq 0$ show that for any $\lambda\in\mathbb{R}$,*

$$\mathbb{P}\left(\sup_{t\in[0,S+T]} X_t\leq\lambda\right) \geq \mathbb{P}\left(\sup_{t\in[0,S]} X_t\leq\lambda\right)\mathbb{P}\left(\sup_{t\in[0,T]} X_t\leq\lambda\right)$$

A important use of Theorem 20 is for the function

$$f(x_1,\dots,x_N) = \frac{1}{\lambda}\log\left(\sum_{i=1}^N e^{\lambda x_i}\right). \tag{8.8}$$

This function is suited to the study of maxima of random variables as seen in Lemma 14. We remark that the derivatives of f satisfy that $\sum_{i=1}^N \partial_i f = 1$ and then $\forall j$ fixed, $\sum_{i=1}^N \partial_{i,j} f = 0$. In particular, we can write that

$$\partial_{i,i} f = -\sum_{j:j\neq i}\partial_{i,j} f. \tag{8.9}$$

We use this fact in the following corollary.

Corollary 6 (Slepian 2)**.** *Let $X = (X_1, \ldots, X_N)$ and $Y = (Y_1, \ldots, Y_N)$ be two centered Gaussian vectors in $\mathbb{R}^N$ such that for all i, j one have $\mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right]$ then,*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \leq \mathbb{E}\left[\max_{i=1,\ldots,N} Y_i\right].$$

Furthermore, if $|\mathbb{E}\left[(X_i - X_j)^2\right] - \mathbb{E}\left[(Y_i - Y_j)^2\right]| \leq \varepsilon$ for all i, j then

$$\left|\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] - \mathbb{E}\left[\max_{i=1,\ldots,N} Y_i\right]\right| \leq \sqrt{\varepsilon \log N}.$$

Proof. Consider the function ϕ of the proof of Theorem 20. In the sequel, we denote by $\delta_{i,j} = \Sigma^Y_{i,j} - \Sigma^X_{i,j}$ and we also omit to write the argument $Z(t)$ of f and its derivatives. Using the notation, $v^X_{i,j} = \mathbb{E}\left[(X_i - X_j)^2\right]$ and $v^Y_{i,j} = \mathbb{E}\left[(Y_i - Y_j)^2\right]$, we see that $v^Y_{i,j} - v^X_{i,j} = \delta_{i,i} + \delta_{j,j} - 2\delta_{i,j}$. By (8.7), we have that

$$\begin{aligned}
\phi'(t) &= \frac{1}{2}\sum_{i,j} \mathbb{E}\left[\partial_{i,j} f\right] \delta_{i,j} \\
&= \frac{1}{2}\sum_{i} \mathbb{E}\left[\partial_{i,i} f\right] \delta_{i,i} + \frac{1}{2}\sum_{i}\sum_{j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] \delta_{i,j} \\
&= -\frac{1}{2}\sum_{i}\sum_{j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] \delta_{i,i} + \frac{1}{2}\sum_{i}\sum_{j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] \delta_{i,j} \\
&= \frac{1}{2}\sum_{i,j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] (\delta_{i,j} - \delta_{i,i}) \\
&= \frac{1}{4}\sum_{i,j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] (2\delta_{i,j} - \delta_{i,i} - \delta_{j,j}) = \frac{1}{4}\sum_{i,j:j\neq i} \mathbb{E}\left[\partial_{i,j} f\right] (v^X_{i,j} - v^Y_{i,j}).
\end{aligned}$$

Moreover, the derivative $\partial_{i,j} f$ are negative (for $i \neq j$) and then this last term is positive under the hypothesis of the first part of the corollary. Hence $\mathbb{E}\left[f(Y)\right] \geq \mathbb{E}\left[f(X)\right]$. But as in the proof of Lemma 14, we use the following inequalities

$$\max_{i=1,\ldots,N} x_i \leq \frac{1}{\lambda}\log\left(\sum_{i=1}^N e^{\lambda x_i}\right) \leq \frac{1}{\lambda}\log N + \max_{i=1,\ldots,N} x_i \tag{8.10}$$

to say that $\mathbb{E}\left[\max_i X_i\right] \leq \mathbb{E}\left[f(X)\right] \leq \mathbb{E}\left[f(Y)\right] \leq \frac{1}{\lambda}\log N + \mathbb{E}\left[\max_i Y_i\right]$ and letting $\lambda \to \infty$ gives the first result. For the second part, we see that since $|\partial_{i,j} f| \leq 1/\lambda$ and $|v^X_{i,j} - v^Y_{i,j}| \leq \varepsilon$, we get that $\phi'(t) \leq \lambda\varepsilon/4$ and then

$$|\mathbb{E}\left[f(X)\right] - \mathbb{E}\left[f(Y)\right]| \leq \left|\int_0^1 \phi'(t)dt\right| \leq \frac{\lambda\varepsilon}{4}.$$

Then using (8.10), we show that

$$\left|\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] - \mathbb{E}\left[\max_{i=1,\ldots,N} Y_i\right]\right| \leq \frac{\lambda\varepsilon}{4} + \frac{\log N}{\lambda}$$

and the result follows by optimizing in λ. □

With this new tool, one is able to prove the following result that is due to Sudakov.

Theorem 21. *Let $X_1, \ldots, X_N$ be centered Gaussian random variables. Then,*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \geq \frac{1}{2}\min_{i\neq j}\sqrt{\mathbb{E}\left[(X_i - X_j)^2\right]\log N}.$$

Proof. Let $Z_1, \ldots, Z_N$ be i.i.d. standard Gaussian variables and set $\delta = \min_{i\neq j}(\mathbb{E}\left[(X_i - X_j)^2\right])^{1/2}$ and finally let $Y_i = \delta Z_i/\sqrt{2}$. By definition of the random variables Y_i, $\mathbb{E}\left[(Y_i - Y_j)^2\right] = \delta^2 \leq \mathbb{E}\left[(X_i - X_j)^2\right]$ and so by using Corollary 6, we get that $\delta\mathbb{E}\left[\max Z_i\right] \leq \sqrt{2}\mathbb{E}\left[\max X_i\right]$. We finish by using the consequence (8.6) of Proposition 27. □

The great consequence of Sudakov minoration is that it gives a lower bound on the suprema of a Gaussian process. Indeed, if $\mathcal{T}$ is a space endowed with the pseudo metric $d(t,s)^2 = \mathbb{E}\left[(X_t - X_s)^2\right]$, then for all $0 < \varepsilon \leq D = \mathrm{Diam}(\mathcal{T})$, then for a centered Gaussian process $(X_t)_t$ we have

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] \geq \frac{1}{2}\varepsilon\sqrt{H(\varepsilon,\mathcal{T},d)}. \tag{8.11}$$

This bound is to be compared with the Dudley entropy bound in Theorem 19. The two orders of the bounds differ. Indeed, Dudley's upper bound is proportional to the integral of the curve $\sqrt{H(\varepsilon,\mathcal{T},d)}$ whereas the bound in (8.11) is proportional to area of the smallest rectangle $\varepsilon \times \sqrt{H(\varepsilon,\mathcal{T},d)}$ which can be significantly smaller when the curve $\sqrt{H(\varepsilon,\mathcal{T},d)}$ is very quickly close to the abscisse axis. In general, for Gaussian processes, the Sudakov lower bound is tight whereas the bound of Dudley is not. There exists a better chaining technique called generic chaining after the invention of Talagrand (see [14]) that allows to bridge the gap between the two bounds. This technique will be exposed in Chapter [WRITE THIS].

8.6 Proof of Theorem 18

In order to prove Theorem 18, we give three successive lemmas that use Dudley bound along with symetrization.

Lemma 16. *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 17. *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 $\sup_f|P_nf - Pf| \xrightarrow{\mathbb{P}} 0$. So one can prove that the convergence occurs in L_1 to have the result. By the symmetrization argument prove in Lemma 12 and Lemma 16, we have

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

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,

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

Then, by the dominated convergence theorem (in its probabilistic version given by Corollary 11), $Y_n \xrightarrow{\mathbb{P}} 0$ 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 18. 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 17, 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. □

8.7 Vapnik-Chervonenkis classes

In this section, we introduce the so-called V-C dimension invented by Vapnik and Chervonenkis [17]. We also refer to [5] for the definitions and combinatorial properties of the notions defined below. The notion of V-C dimension raised in the study of bounds for empirical processes of the form

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$$

where μ_n is the empirical measure corresponding to μ. As one can expect, the efficiency of the convergence depends deeply on the class of sets $\mathcal{A}$. This context is simpler that the context of the beginning of the chapter but is informative in general.

Definition 9. *Let $x_1, \ldots, x_n \in \mathbb{R}^k$ be fixed points. We define the **trace** of the set $\mathcal{A}$ over the collection $x_1, \ldots, x_n$ as*

$$\mathcal{A}(x_1^n) = \{(\mathbb{1}_{x_1 \in A}, \ldots, \mathbb{1}_{x_n \in A}) \in \{0,1\}^n \ : A \in \mathcal{A}\}.$$

*The **shatter coefficient** is given by*

$$S_{\mathcal{A}}(n) = \max_{(x_1, \ldots, x_n) \in \mathbb{R}^k} |\mathcal{A}(x_1^n)|.$$

*By definition, $S_{\mathcal{A}}(n) \leq 2^n$. The last n such that the inequality is an equality is called the **dimension Vapnik and Chervonenkis** of $\mathcal{A}$, this is*

$$V = \sup\{n \in \mathbb{N}^* \ : S_{\mathcal{A}}(n) = 2^n\}.$$

If $\forall n \in \mathbb{N}^*$ the shatter coefficient equals 2^n we define $V = \infty$.

Some examples of calculation of VC dimension

In this small section, we describe some examples of contexts where it is possible to show that the class admits a finite VC dimension.

- Let $\mathcal{A} = \{(-\infty, t] : t \in \mathbb{R}\}$. By the sake of simplicity we assume that the x_i are ordered $x_1 < x_2 < \cdots < x_n$. Then the elements of $\mathcal{A}(x_1^n)$ are of the form $(1, \ldots, 1, 0, \ldots, 0)$. There are exactly $n+1$ possible choices and then $S_{\mathcal{A}}(n) = n+1$. Since for $n = 1$, we have that $n+1 = 2^n$ but for any $n > 1$, $n+1 < 2^n$ we have shown that the class $\mathcal{A}$ has VC dimension equal to 1.

- Let $\mathcal{A} = \{[a, b] : a, b \in \mathbb{R}\}$. Assuming again the x_i ordered, the elements of $\mathcal{A}(x_1^n)$ are of the form $(0, 0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$. There are $\binom{n+2}{2} - n = (n^2+n+2)/2$ such elements. Then $S_{\mathcal{A}}(n) = (n^2+n+2)/2$. But for $n = 1, 2$ we have that $S_{\mathcal{A}}(n) = 2^n$ but for $n > 2$ we have $S_{\mathcal{A}}(n) < 2^n$. In this case, $V = 2$.

- Let $\mathcal{A} = \{\otimes_{i=1}^d (-\infty, t_i] : t_i \in \mathbb{R}\}$ for $d \geq 2$. By a study similar to the first case, one can show that $S_{\mathcal{A}}(n) \leq (n+1)^d$ and this allows to bound the value of V.

- Let $\mathcal{A} = \{\{x : \theta^T x > y\} : \theta \in \mathbb{R}^d, y \in \mathbb{R}\}$ the class of half space of $\mathbb{R}^d$. Then we have that $S_{\mathcal{A}}(n) \leq 2^d \binom{n}{d}$ and a few more calculation give that $V \leq d+1$.

Exercice 24. *Show that in the last example, $S_{\mathcal{A}}(n) \leq 2^d \binom{n}{d}$. Hint: The hyperplanes are completely defined by d points and the rest lie one one or the other side.*

8.7.1 Sauer's Lemma

Sauer's Lemma is a result that allows to show that a class $\mathcal{A}$ that has a finite V-C dimension has shatter coefficients that grow at a polynomial speed in n.

Lemma 18 (Sauer)**.** *Let $\mathcal{A}$ be a class of finite V-C dimension V. Then, for all $n \geq 1$,*

$$S_{\mathcal{A}}(n) \leq \sum_{i=1}^{V} \binom{n}{i}.$$

Proof. We need the following definition. We say that a set $B \subset \{0,1\}^n$ **shatters** a set $S = \{s_1, \ldots, s_m\} \subset \{1, \ldots, n\}$ if the restriction of B to the components $s_1, \ldots, s_m$ is the full hypercube, that is

$$B_S := \{(b_{s_1}, \ldots, b_{s_m}) \ : (b_1, \ldots, b_n) \in B\} = \{0,1\}^m.$$

We define the transformation

$$\begin{aligned} \Psi_1 \colon \mathcal{P}(\{0,1\}^n) &\to \mathcal{P}(\{0,1\}^n) \\ B &\mapsto \Psi_1(B) = \{\bar{b} : b \in B\} \end{aligned}$$

where $\forall b = (b_1, \ldots, b_n) \in B$ we define $\bar{b}$ by:

1. If b is such that $b_1 = 1$, then $\bar{b} = (0, b_2, \ldots, b_n)$ if $(0, b_2, \ldots, b_n) \notin B$ and $\bar{b} = b$ otherwise.

2. If b is such that $b_1 = 0$, then $\bar{b} = b$.

- Fact 1: For $B \subset \{0,1\}^n$ and $S = \{s_1, \ldots, s_m\}$, we have that $|\Psi_1(B)| = |B|$ and B shatters S if and only if $\Psi_1(B)$ shatters S.

First of all, it is clear that Ψ_1 is injective which imply that $|\Psi_1(B)| = |B|$. Now assume that B shatters S. If $1 \notin S$, then $(\Psi_1(B))_S = B_S$ and this case is obvious. In the other case, since $B_S = \{0,1\}^m$ there is no room for case 1 to modify b. Then, we always have $\bar{b} = b$ and again $(\Psi_1(B))_S = B_S$.

- Fact 2: Define $\Psi_2, \ldots, \Psi_n$ analog transformations on the coordinates $2, \ldots, n$ and let $B_0 \subset \{0,1\}^n$ and $B_n = \Psi_n \circ \cdots \circ \Psi_1(B_0)$. Assume that any set S of m indexes with $m > V$ are not shattered by B_0. For an element v of the final set B_n, we define
$$T_v = \{b \in \{0,1\}^n : \ b_i \leq v_i\}.$$
Then $T_v \subset B_n$ and v does not have more than V ones.

Indeed, we have seen that $(1, v_2, \ldots, v_n) \in B_1 \Rightarrow (0, v_2, \ldots, v_n) \in B_1$. The following transformations may change the vector $v_2, \ldots, v_n$ but in both $(1, v_2, \ldots, v_n)$ and $(0, v_2, \ldots, v_n)$ in the same way, then $(1, v_2, \ldots, v_n) \in B_n \Rightarrow (0, v_2, \ldots, v_n) \in B_n$. Similarly, $\forall i$, $(v_1, \ldots, 1, \ldots, v_n) \in B_i \Rightarrow (v_1, \ldots, 0, \ldots, v_n) \in B_i$, that transfers to B_n too. This implies that every vector with entries $b_i \leq v_i$ are in B_n or equivalently, $T_v \subset B_n$.
Since B_0 does not shatter S, B_n either (by fact 1). Now, assume that v has $m > V$ ones and take $S = \{i : v_i = 1\}$. S is a set such that $\{0,1\}^m = (T_v)_S \subset (B_n)_S$ and then B_n shatters S but this is absurd then v cannot have more that V ones.

- Fact 3: Let $T = \cup_{v \in \mathcal{V}} T_v$ where $\mathcal{V}$ is the set of all vectors with no more that V ones. Then $|B_n| \subset T$ and $|T| = \sum_{i=0}^{V} \binom{n}{i}$.

Since by fact 2, there is no vectors with more than V ones in B_n, we directly have $B_n \subset T$. T can be rewritten as the disjoint union of the sets of vectors with exactly i ones. This gives, $|T| = \sum_{i=0}^{V} \binom{n}{i}$.

- Fact 4: Conclusion

The inequality is trivial for $n \leq V$ since the sum equals 2^n in that case. For the case $n > V$, let $x_1, \ldots, x_n \in \mathbb{R}^k$ such that $B_0 = \mathcal{A}(x_1^n)$ and $S_{\mathcal{A}}(n) = |B_0|$ (which exists by definition of $S_{\mathcal{A}}(n)$). If one set of indexes S with $m > V$ would be shattered by B_0 that would mean that $S_{\mathcal{A}}(m) = 2^m$ which is absurd! Then, we are in the case of fact 2 and

$$S_{\mathcal{A}}(n) = |B_0| = |B_n| \leq \sum_{i=0}^{V} \binom{n}{i}.$$

□

As a direct consequence, we show that the shatter coefficient cannot grow exponentially fast when the V-C dimension is finite as stated in the following result.

Corollary 7. *Let $\mathcal{A}$ be a class of finite V-C dimension V. Then, for any n,*

$$S_{\mathcal{A}}(n) \leq (n+1)^V$$

and in the special case $n \geq V$ we have,

$$S_{\mathcal{A}}(n) \leq \left(\frac{ne}{V}\right)^V.$$

Proof. We have for all n,

$$\sum_{i=0}^{V} \binom{n}{i} \leq \sum_{i=0}^{V} \frac{n^i}{i!} \leq \sum_{i=0}^{V} \frac{n^i V!}{i!(V-i)!} \leq \sum_{i=0}^{V} n^i \binom{V}{i} \leq (n+1)^V.$$

If $n \geq V$ so $V/n \leq 1$ and

$$\left(\frac{V}{n}\right)^V \sum_{i=0}^{V} \binom{n}{i} \leq \sum_{i=0}^{V} \left(\frac{V}{n}\right)^i \binom{n}{i} \leq \sum_{i=0}^{n} \left(\frac{V}{n}\right)^i \binom{n}{i} \leq \left(1 + \frac{V}{n}\right)^n \leq e^V.$$

□

8.7.2 Entropy on the hypercube

We define the **Hamming distance** ρ on $\{0,1\}^n$ between two elements $b, c \in \{0,1\}^n$ by,

$$\rho(b,c) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{b_i \neq c_i}}.$$

We define a probability measure on a class of sets $\mathcal{A}$ relatively to a measure Q by

$$d_Q(A,B) = (Q(A \triangle B))^{1/2} = \|\mathbb{1}_A - \mathbb{1}_B\|_{2,Q}.$$

The following theorems deal with the fact that V-C dimension bounds the entropy of $\mathcal{A}$ endowed with d_Q and $\mathcal{A}(x_1^n)$ endowed with ρ.

Theorem 22. *Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^k$ of V-C dimension $V < \infty$. Then, for each $x_1, \ldots, x_n \in \mathbb{R}^d$ y $0 \leq \varepsilon \leq 1$,*

$$H(\varepsilon, \mathcal{A}(x_1^n), \rho) \leq \frac{V}{1 - 1/e} \log \frac{4e}{\varepsilon^2}$$

Proof. Let C_ε be a ε-net of the class $\mathcal{A}(x_1^n)$. We denote by $c^{(1)}, \dots, c^{(M)}$ the elements of the ε-net C_ε. For two different points $c^{(i)}$ and $c^{(j)}$, we introduce

$$A_{i,j} = \{t \in \{1, \dots, k\} : c_t^{(i)} \neq c_t^{(j)}\}$$

Let $Y_1, \dots, Y_K$ be K independent uniform random variables on $\{1, \dots, k\}$. Then by the fact that $c^{(i)}$ and $c^{(j)}$ are at distance at least ε from each other, we have that $\mathbb{P}(Y_m \in A_{i,j}) = |A_{i,j}|/k \geq \varepsilon^2 k/k = \varepsilon^2$. Then

$$\begin{aligned}\mathbb{P}(\forall i, j \leq M : i \neq j, \text{ at least one } Y_m \text{ belongs to } A_{i,j}) &= 1 - \mathbb{P}(\exists i, j \leq M : i \neq j, \forall m,\ Y_m \notin A_{i,j}) \\ &\geq 1 - M^2 \mathbb{P}(\forall m,\ Y_m \notin A_{1,2}) \\ &= 1 - M^2 \prod_{m=1}^{K} \mathbb{P}(Y_m \notin A_{1,2}) \\ &\geq 1 - M^2(1-\varepsilon^2)^K \geq 1 - M^2 e^{-K\varepsilon^2}.\end{aligned}$$

By taking $K = \lfloor 2 \log M/\varepsilon^2 \rfloor + 1$, we see that the above probability is positive. In other words, there is a realization of the random variables $Y_1(\omega) = y_1, \dots, Y_K(\omega) = y_K$ that satisfy the event $\bigcap_{i \neq j} \bigcup_m \{y_m \in A_{i,j}\}$. Since we have shown that for any i and j, there exists at least one coordinate y_m for which $c_{y_m}^{(i)} \neq c_{y_m}^{(j)}$, any two elements of $(C_\varepsilon)_{y_1, \dots, y_K}$ (that is defined as the restriction of the elements of C_ε to the coordinates $y_1, \dots, y_K$) are also distinct. Then $|(C_\varepsilon)_{y_1, \dots, y_K}| = |C_\varepsilon| = M$. Without loosing in generality, we assume that $|\{y_1, \dots, y_K\}| = K$ (or that the y_m are all distinct) since the previous property remains true if one add extra coordinates. By Sauer lemma, for $K \geq V$,

$$M = |(C_\varepsilon)_{y_1, \dots, y_K}| \leq |\mathcal{A}(y_1^K)| \leq S_{\mathcal{A}}(K) \leq \left(\frac{eK}{V}\right)^V.$$

Now, if $\log M \geq V$ then $K \geq V$ and we also have that

$$\log M \leq V \log\left(\frac{eK}{V}\right) \leq V \log\left(\frac{4e \log M}{V \varepsilon^2}\right) = V\left(\log \frac{4e}{\varepsilon^2} + \log \frac{\log M}{V}\right) \leq V \log \frac{4e}{\varepsilon^2} + \frac{1}{e} \log M,$$

from which we directly deduce that $\log M \leq \frac{V}{1-1/e} \log \frac{4e}{\varepsilon^2}$. Otherwise, $\log M < V$ and this last fact is trivially verified. By Proposition 25, C_ε is a covering of the set $\mathcal{A}(x_1^n)$ and then $H(\varepsilon, \mathcal{A}(x_1^n), \rho) \leq \log M$ □

Theorem 23. *Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^k$ of V-C dimension $V < \infty$. Let Q be a probability measure on $\mathbb{R}^k$. Then, for each $0 \leq \varepsilon \leq 1$,*

$$H(\varepsilon, \mathcal{A}, d_Q) \leq \frac{V}{1-1/e} \log \frac{4e}{\varepsilon^2}$$

Proof. We proceed as in the proof of Theorem 22. Let C_ε be a ε-net of the set $\mathcal{A}$. The denote $A^{(1)}, \dots, A^{(M)}$ the points (that are events) of C_ε. Let $Y_1, \dots, Y_K$ be i.i.d. random variables of law Q. Then $\mathbb{P}\left(Y_m \in A^{(i)} \triangle A^{(j)}\right) \geq \varepsilon^2$ be definition of the ε-net. We finish the proof by the exact same arguments as in the proof of Theorem 22. □

8.7.3 V-C classes for functions

In this section, we link the entropy on spaces of functions with a notion of V-C dimension for a class of sets related to those functions. For a real valued function $f : \mathcal{X} \to \mathbb{R}$, we define its **subgraph** as the set

$$\mathrm{SubGr}(f) = \{(x, y) \in \mathcal{X} \times \mathbb{R} : y < f(x)\}.$$

By an obvious use of Fubini theorem, for a probability measure Q on $\mathcal{X}$ and a non-negative function f,

$$\int_{\mathcal{X}} f(x) dQ(x) = \int_{\mathcal{X}} \int_0^{f(x)} 1 dy dQ(x) = \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{0 \leq y \leq f(x)} d(Q \times \lambda)(x, y)$$

where λ holds for Lebesgue measure on $\mathbb{R}$. It is then clear that for two functions f and g for which we denote by G_f and G_g their respective subgraphs,

$$\begin{aligned}\int_{\mathcal{X}} |f(x) - g(x)| dQ(x) &= \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{0 \leq y \leq g(x) - f(x)} d(Q \times \lambda)(x, y) + \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{0 \leq y \leq f(x) - g(x)} d(Q \times \lambda)(x, y) \\ &= \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{f(x) \leq y \leq g(x)} d(Q \times \lambda)(x, y) + \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{g(x) \leq y \leq f(x)} d(Q \times \lambda)(x, y) \\ &= \int_{\mathcal{X} \times \mathbb{R}} \mathbb{1}_{G_f \triangle G_g}(x, y) d(Q \times \lambda)(x, y) \\ &= (Q \times \lambda)(G_f \triangle G_g) \qquad (8.12)\end{aligned}$$

where we used on the second step the fact that the Lebesgue measure is translation invariant. This motivates the following definition.

Definition 10. *Let $\mathcal{F}$ be a class of real valued functions. We define the* **VC dimension** *of the class $\mathcal{F}$ as the VC dimension of the class*

$$\mathcal{G} = \{\mathrm{SubGr}(f) : f \in \mathcal{F}\}.$$

We say that a class $\mathcal{F}$ is a **VC class** *if its VC dimension is finite.*

Example 8. *A somewhat obvious consideration is that the set of functions*

$$\mathcal{F} = \{\mathbb{1}_A : A \in \mathcal{A}\},$$

where $\mathcal{A}$ is a VC class, forms a VC class (in the sense of Definition 10) of enveloppe given by $F = 1$. This is particularly well suited for density/measure estimation over a class of events that form a VC class. The following theorems show that the entropy of such a class of indicator functions is therefore, for a constant $A > 0$

$$H_1(\varepsilon, \mathcal{F}, P_n) \leq A \log\left(\frac{1}{\varepsilon}\right), \quad \forall \varepsilon > 0.$$

In particular, the set $\mathcal{F}$ is P-Glivenko-Cantelli by the use of Theorem 18.

The tools developed for VC classes of sets can be used in this context to bound the entropy of a set of functions $\mathcal{F}$ by a explicit formula that only depends on the VC dimension, the radius of the balls of the covering ε and the L_r norm of the enveloppe. It is remarkable that the following upper bound only depends on the subjacent probability measure Q through $\|F\|_{r,Q}$, the $L_r(Q)$ norm of the enveloppe.

Theorem 24. *Let $\mathcal{F}$ be a class of VC dimension V and such that for a probability measure Q, the enveloppe function $F \in L_1(Q)$ then for any $\varepsilon \in (0, 1)$,*

$$H_1(\varepsilon\|F\|_{1,Q}, \mathcal{F}, Q) \leq \frac{V}{1 - 1/e} \log \frac{8e}{\varepsilon}.$$

If, in addition, $F \in L_r(Q)$ then for any $\varepsilon \in (0, 1)$, one have that

$$H_r(\varepsilon\|F\|_{r,Q}, \mathcal{F}, Q) \leq \frac{V}{1 - 1/e} \log \frac{8e}{(\varepsilon/2)^r}.$$

Proof. We begin with the case $r = 1$ from which we will deduce the general case. As seen in Equation (8.12), $Q|f - g| = (Q \times \lambda)(G_f \triangle G_g)$. We consider the probability P that is the measure $Q \times \lambda$ conditioned to the set $\{(x, y) : |y| \leq F(x)\}$. Then $P = (Q \times \lambda)/2\|F\|_{1,Q}$. By using Theorem 23, we get that

$$H_1(2\varepsilon\|F\|_{1,Q}, \mathcal{F}, Q) = H(2\varepsilon\|F\|_{1,Q}, \mathcal{G}, Q \times \lambda) = H(\sqrt{\varepsilon}, \mathcal{G}, d_P) \leq \frac{V}{1 - 1/e} \log \frac{4e}{\varepsilon}$$

where $\mathcal{G}$ is the set of the subgraphs of the functions in $\mathcal{F}$. This gives the result for $r = 1$ by replacing ε by $\varepsilon/2$ in the previous chain of inequalities. For $r > 1$, by defining the probability measure R such that $dR/dQ = F^{r-1}/QF^{r-1}$ we can write

$$Q|f - g|^r \leq Q|f - g|(2F)^{r-1} = 2^{r-1} QF^{r-1} \times R|f - g|$$

so that $\|f - g\|_{r,Q} \leq 2(QF^{r-1})^{1/r} \times \|f - g\|_{1,R}^{1/r}$. By direct comparison of the entropies for different distances, we get

$$H(2\varepsilon\|F\|_{r,Q}, \mathcal{F}, \|\cdot\|_{r,Q}) \leq H\left(\varepsilon^r \frac{\|F\|_{r,Q}^r}{QF^{r-1}}, \mathcal{F}, \|\cdot\|_{1,R}\right) = H_1(\varepsilon^r\|F\|_{1,R}, \mathcal{F}, R) \leq \frac{V}{1 - 1/e} \log \frac{8e}{\varepsilon^r}.$$

□

As a direct consequence of the preceding result, one can show that a VC class have a finite entropy for the L_1 norm as long as the enveloppe is integrable. Indeed, one can use Theorem 24 for any probability measure $Q = P_n$ and since the bound is completely uniform, one obtain

Corollary 8. *Let $\mathcal{F}$ be a VC class such that the enveloppe F belongs to $L_1(P)$, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

8.8 P-Glivenko-Cantelli classes through convexity

It is remarkable that the property of being P-Glivenko-Cantelli behaves well with the convexification of a set $\mathcal{F}$.

Proposition 28. *Let $\mathcal{F}$ be a P-Glivenko-Cantelli class then its convex hull $Conv(\mathcal{F})$ defined by*

$$Conv(\mathcal{F}) = \left\{ \sum_{j=1}^{p} \theta_j f_j : p \in \mathbb{N}^*, \theta_j \geq 0 \text{ and } \sum_{j=1}^{p} \theta_j = 1 \right\} \tag{8.13}$$

is also P-Glivenko-Cantelli.

Proof. Let $p \geq 1$, let $(f_j)_{j=1,\dots,p} \in \mathcal{F}$ and $\sum_j \theta_j = 1$, then

$$\begin{aligned}\left| \int \sum_{j=1}^{p} \theta_j f_j d(P_n - P) \right| &= \left| \sum_{j=1}^{p} \theta_j \int f_j d(P_n - P) \right| \\ &\leq \left(\sum_{j=1}^{p} \theta_j \right) \max_{j=1,\dots,p} \left| \int f_j d(P_n - P) \right| \\ &\leq \sup_{f \in \mathcal{F}} \left| \int f d(P_n - P) \right|\end{aligned}$$

Then $\sup_{f \in \mathrm{Conv}(\mathcal{F})} |P_n f - Pf| = \sup_{f \in \mathcal{F}} |P_n f - Pf|$ and since $\mathcal{F}$ is P-Glivenko Cantelli, so is Conv($\mathcal{F}$). □

This convenient fact is very useful in practice since one can prove that the extreme points of $\mathcal{F}$ form a P-Glivenko Cantelli class to get that the entire class $\mathcal{F}$ is also P-Glivenko Cantelli.

8.9 Dudley entropy bound and Orlicz norm

8.9.1 A global bound

The ideas behind the proof of Dudley entropy bound are possible to adapt to prove the concentration of the supremum over a class of random variables with sub-Gaussian increments. These fact will be useful to use the so-called peeling device.

Theorem 25. *Let $(\mathcal{T}, d)$ be a metric space and let $(X_t)_{t \in \mathcal{T}}$ be a random process such that for all $s, t \in \mathcal{T}$,*

$$\|X_s - X_t\|_{\psi_2} \leq d(s,t)$$

where $\|\cdot\|_{\psi_2}$ is the Orlicz norm of section 7.3. Then, for all $t_0 \in \mathcal{T}$,

$$\left\| \sup_{t \in \mathcal{T}} |X_t - X_{t_0}| \right\|_{\psi_2} \leq 12 \int_0^{\delta/2} \sqrt{H(\varepsilon, \mathcal{T}, d)} d\varepsilon$$

where $\delta = \sup_{t \in \mathcal{T}} d(t, t_0)$.

In particular, we see that the suprema of a sub-Gaussian process is a sub-Gaussian random variable. The result of theorem 25 is global since it is true for all range of deviation. Indeed, there exist constants $K, C > 0$ such that

$$\mathbb{P}\left(|\sup_{t \in \mathcal{T}} |X_t - X_{t_0}| \geq a \right) \leq K \exp\left(-C \frac{a^2}{\int_0^{\delta/2} \sqrt{H(\varepsilon, \mathcal{T}, d)} d\varepsilon} \right), \quad \forall a > 0. \tag{8.14}$$

The entropy bound is independent of a so the name of global bound. In some range of applications, it can be sufficient to have a bound that is valid for a range of a that is limited. Such bounds do not require to refine the covering until that the radius of the balls that cover T tends to 0. This idea is at the origin of the *stopped chaining* and of the *peeling argument*.

Chapter 9

Deviation bounds and peeling device

9.1 A local bound through uniform discretization

In this section we investigate a special case where the discretization given by the chaining technique can be refined. We assume that the process $(X_t)_{t\in\mathcal{T}}$ is sub-Gaussian and can be discretized uniformly by a set $\mathcal{S}_\delta$ such that

$$\sup_{t\in\mathcal{T}} \inf_{s\in S_\delta} |X_t - X_s| \leq \delta. \tag{9.1}$$

Under this assumption, one will be able to bound the deviation of order δ of the suprema of the process $(X_t)_t$.

The empirical process Going back to the empirical process $((P_n - P_n')f)_{f\in\mathcal{F}}$, it is possible to verify the previous condition using the empirical distance. If we assume that the entropy $H_1(\delta, \mathcal{F}, P_n - P_n')$ is finite, there exists a covering of the set $\mathcal{F}$ with $\mathcal{C}_\delta = (g_i)_i$ as the set of centers. In this case, the condition is

$$\sup_{f\in\mathcal{F}} \inf_{g_i\in\mathcal{C}_\delta} |(P_n - P_n')f - (P_n - P_n')g_i| \leq \sup_{f\in\mathcal{F}} \inf_{g_i\in\mathcal{C}_\delta} \|f - g_i\|_{1,P_n - P_n'} \leq \delta.$$

In this case we have the following result.

Proposition 29. *Let $(X_t)_{t\in\mathcal{T}}$ be a stochastic process on the space $\mathcal{T}$. We assume that there exists a constant $V > 0$ such that every random variable X_t verifies $\|X_t\|_{\psi_p}^p \leq V$ for $\psi_p : x \to e^{x^p} - 1$ and $p > 0$. Assume moreover that the condition (9.1) is verified for a value $\delta > 0$. Then, there exists a universal constant $K > 0$,*

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}} |X_t| \geq 2\delta\right) \leq 2\exp\left(-K\frac{\delta^p}{V\psi_p^{-1}(|S_\delta|)}\right).$$

Proof. Since the functions ψ_p satisfy in particular the conditions of Proposition 21, the result holds by noticing that

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}} |X_t| \geq 2\delta\right) \leq \mathbb{P}\left(\max_{s\in\mathcal{S}_\delta} |X_s| \geq \delta\right).$$

□

One can also derive the direct bound that follows from a similar idea as in (7.6) and obtain

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}} |X_t| \geq 2\delta\right) \leq \exp\left(-K\frac{\delta^p}{V} + \log(|S_\delta|)\right).$$

A natural example is the case of empirical processes again where the uniform discretization is given by the centers of the ball in the empirical covering. The set $\mathcal{S}_\delta$ is the centers of the balls of radius δ that cover (for the empirical distance given by the norm $\|\cdot\|_{1,P_n-P_n'}$) and $|\mathcal{S}_\delta| = \mathcal{N}_1(\varepsilon, \mathcal{F}, P_n - P_n')$.

9.2 The peeling device

In this section, we assume that the functions in $\mathcal{F}$ are such that

$$\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M.$$

The idea behind the peeling device is to cut the set $\mathcal{F}$ by levels that have a compared value for a given criteria. More precisely, we assume given a function $\tau : \mathcal{F} \to [\Delta, R)$ such that $\Delta > 0$ and R is any positive real or $+\infty$. Let $(m_s)_{s=0}^S$ be an increasing sequence of numbers such that $m_0 = \Delta$ and $m_S = R$. In the special case of $R = +\infty$, we impose $S = +\infty$ and the condition $m_S = R$ takes the form of $m_s \to +\infty$ when $s \to +\infty$. We say that $(\mathcal{F}_s)_{s=1}^S$ is a **peeling** of $\mathcal{F}$ if

$$\mathcal{F} = \bigcup_{s=1}^{S} \mathcal{F}_s$$

and $\mathcal{F}_s = \{f \in \mathcal{F} : m_{s-1} \le \tau(f) < m_s\}$. The key point of the peeling idea is the following result.

Lemma 19. *Let $(X_f)_{f\in\mathcal{F}}$ be a stochastic process defined on the set of functions $\mathcal{F}$. Then for any $a > 0$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{|X_f|}{\tau(f)} \ge a\right) \le \sum_{s=1}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F}:\tau(f)<m_s} |X_f| \ge a m_{s-1}\right).$$

Proof. The proof consists in a simple break of the supremum.

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{|X_f|}{\tau(f)} \ge a\right) &\le \mathbb{P}\left(\sup_{s=1,\dots,S}\sup_{f\in\mathcal{F}_s} \frac{|X_f|}{\tau(f)} \ge a\right)\\
&\le \sum_{s=1}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F}_s} \frac{|X_f|}{\tau(f)} \ge a\right)\\
&\le \sum_{s=1}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F}_s} \frac{|X_f|}{m_{s-1}} \ge a\right)\\
&\le \sum_{s=1}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F}:\tau(f)<m_s} |X_f| \ge a m_{s-1}\right),
\end{aligned}$$

where we used the definition of the set $\mathcal{F}_s$ in the last two inequalities. □

The peeling device is particularly useful to control the empirical process deviation using simple bounds as Proposition 29. The bounds that we derive now take the form of deviation bounds on the ratio $\sup(P_n - P)f/\|f\|$. The following result is key to detail the assumptions of Donsker theorems that will be studied in the following chapter.

The Rademacher process We detail the peeling device for the process that one has encountered a few times, the process $(\nu_f)_{f\in\mathcal{F}} = (n^{-1}\sum_i \varepsilon_i f(X_i))_{f\in\mathcal{F}}$. For this purpose, we use the peeling device for the specific increasing sequence $m_s^2 - m_0^2 \ge s$ and $m_0 > 0$ and the criteria function $\tau(f) = \|f\|_2$. Since the process is sub-Gaussian conditionally to the variables X_i, we write that

$$\mathbb{P}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{\nu_f}{\|f\|_2} \ge 2a\right) \le \sum_{s=1}^{\infty} \mathbb{P}_\varepsilon\left(\sup_{f\in\mathcal{F}:\|f\|_2<m_s} \nu_f \ge 2a m_{s-1}\right) \le \sum_{s=1}^{\infty} 2\exp\left(-n\frac{a^2 m_{s-1}^2}{8M^2} + H_2(a m_{s-1}, \mathcal{F}, P_n)\right)$$

where we used Hoeffding inequality for the empirical process ν_f and we used the fact that $\|f\|_\infty \le M$. Finally one have that

$$\begin{aligned}
\mathbb{P}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{\nu_f}{\|f\|_2} \ge 2a\right) &\le 2\exp(H_2(a, \mathcal{F}, P_n)) \times \sum_{s=0}^{\infty} \exp(-a^2 m_s^2 n)\\
&\le 2\exp(H_2(a, \mathcal{F}, P_n)) \times \frac{\exp(-a^2 m_0 n)}{1 - \exp(-a^2 n)}
\end{aligned}$$

Chapter 10

Uniform Central Limit Theorems

In this chapter, we derive central limit theorems that will be valid for empirical processes. Those can also be called uniform central limit theorems. In all this chapter we will be interested in studying the limiting behavior of the process

$$Z_n = \left\{ Z_n(f) = \sqrt{n}(P_n f - Pf) : \; f \in \mathcal{F} \right\}.$$

We also assume that a specific element f_0 is of particular interest and define,

$$\mathcal{F}(\delta) = \{ f \in \mathcal{F} : \; \|f - f_0\| \leq \delta \}.$$

We have to give a precise meaning to the convergence in distribution of the process Z_n. In the notions of convergence given in Chapter 2, we had to have the notion of distance on the space of the values of the random variables. The problem is similar to the measurability of the supremum of a random process that we faced in the beginning of Chapter 8.

Weak convergence of random processes To speak about weak convergence of the random processes, we first have to ensure that the object 'random process' is an element of a metric space. We successively increase in difficulty/generality along with the cases below. We denote by $Z : f \in \mathcal{F} \to \mathbb{R}$ the random process that associates to any element $f \in \mathcal{F}$ a real random number denoted by $Z(f)$. To be able to use distances, we will assume some structure on the realizations (also called trajectories) of the random process.

1. When the trajectories of the process Z are bounded on $\mathcal{F}$, we can use the supremum norm $\| \cdot \|_\infty$ defined as

   $$\|Z\|_\infty = \sup_{f \in \mathcal{F}} |Z(f)|.$$

   We denote by $\ell^\infty(\mathcal{F})$, the space of functions defined on $\mathcal{F}$ that have a finite infinite norm. Obviously, the space $(\ell^\infty(\mathcal{F}), \| \cdot \|_\infty)$ is a metric space.

2. If the space $\mathcal{F}$ can be endowed with a measure structure that makes of $\mathcal{F}, \Sigma$ a measurable space, then one can define the $L_p(\mathcal{F})$ spaces for $1 \leq p < \infty$. The spaces $(L_p(\mathcal{F}), \| \cdot \|_p)$ where the norm is the Lebesgues norm are metric spaces. Then if the trajectories of the process are elements of a same $L_p(\mathcal{F})$, they belong to a metric space.

3. When the trajectories of the process Z are almost surely bounded and the space $\mathcal{F}$ is measurable, one can use the

   $$\|Z\|_{essup} = \inf\{C \geq 0 : |Z| \leq C \text{ almost surely.}\}$$

   The space of trajectories almost surely bounded is denoted by $L_\infty(\mathcal{F})$ and when associated with the norm $\| \cdot \|_{essup}$, it is a metric space.

4. If the space $\mathcal{F}$ can be defined as $\mathcal{F} = \bigcup_{i=1}^\infty \mathcal{F}_i$ and denote $\ell^\infty((\mathcal{F}_i)_i)$ the space of functions $f : \mathcal{F} \to \mathbb{R}$ such that all the restrictions $f_{|\mathcal{F}_i}$ are bounded. We also denote by $\|f\|_{\mathcal{F}_i,\infty}$ the values $\|f_{|\mathcal{F}_i}\|_\infty$. Then one can define the norm

   $$\|f\| = \sum_{i=1}^\infty (\|f\|_{\mathcal{F}_i,\infty} \wedge 1) 2^{-i},$$

   that makes the space $(\ell^\infty((\mathcal{F}_i)_i), \| \cdot \|)$ a metric space.
   For example, if one takes the spaces $\mathcal{F}_i$ to be the intervals $[-i, i]$, this metric is the metric of the uniform convergence on every compact of $\mathbb{R}$.

In the following, we do not detail in which of these context we assume to be in but only represent the underlying metric space as $(\mathbb{D}, \|\cdot\|)$. The metric structure allows us to talk about continuous functions and Lipschitz functions on the space $\mathbb{D}$. For example, a Lipschitz transformation of the process Z is $h(Z)$ where the function $h : \mathbb{D} \to \mathbb{R}$ is such that $\forall x, y \in \mathbb{D}$, $|h(x) - h(y)| \leq \lambda \|x - y\|$.

Definition 11 (Convergence in distribution in $\mathbb{D}$)**.** *We say that a sequence of random processes $(Z_n)_n$ of values in a metric space $(\mathbb{D}, \|\cdot\|)$ converges weakly to a random process Z iff for any Lipschitz function $h : \mathbb{D} \to \mathbb{R}$*

$$\mathbb{E}\left[h(Z_n)\right] \to \mathbb{E}\left[h(Z)\right]. \tag{10.1}$$

We still denote $Z_n \xrightarrow{(d)} Z$ for this convergence. Obviously, the notion of convergence is dependent on the topology given by the metric space $(\mathbb{D}, \|\cdot\|)$.

This definition imposes in particular that every finite dimensional marginal $(Z_n(f_1), \ldots, Z_n(f_k))$ converges in distribution to the finite dimensional marginal $(Z(f_1), \ldots, Z(f_k))$.

Exercice 25. *Prove that the last sentence holds true.*

For the particular case of empirical processes, we define the notion that replaces the convergence in distribution toward a Gaussian random variable. We recall that a **Gaussian process** G on a set $\mathcal{T}$ is a collection of random variables $G = (G_t)_{t \in \mathcal{T}}$ such that for any finite set $(t_1, \ldots, t_k)$ of elements of $\mathcal{T}$, the vector $(G_{t_1}, \ldots, G_{t_k})$ is Gaussian.

Definition 12 (P-Donsker)**.** *For a sample $X_1, \ldots, X_n$ of common law P, we define the normalized empirical process by*

$$Z_n = \left\{ Z_n(f) = \sqrt{n}(P_n f - Pf) : \ f \in \mathcal{F} \right\}$$

*where $\mathcal{F} \subset L_2(\mathbb{R})$. We say that the process Z_n is P-**Donsker** if $Z_n \xrightarrow{(d)} G$ where $G = \{G(f) : \ f \in \mathcal{F}\}$ is the unique centered Gaussian process such that* $\operatorname{Cov}(G(f), G(f')) = Pff' - PfPf'$ *for any two functions $f, f' \in \mathcal{F}$.*

The study of the first chapters showed that the convergence in distribution of Z_n cannot occur towards a random process that would not be a Gaussian process. Indeed, by the TCL theorem (in Theorem 6) any marginals $(Z_n(f_1), \ldots, Z_n(f_k))$ converges in distribution to a Gaussian vector.

10.1 A fundamental Lemma towards P-Donsker classes

We derive a fundamental sufficient condition for a class $\mathcal{F}$ to be P-Donsker in the case where the trajectories of the empirical process are a.s. bounded.

Lemma 20. *Let $\mathcal{F}$ be a class of functions included in $L_2(P)$ that is totally bounded. Let Z_n be a random process such that a.s. its trajectories belong to $\ell_\infty(\mathcal{F})$. We assume that $\forall \eta > 0$, $\exists \delta > 0$,*

$$\limsup_{n \to \infty} \mathbb{P}\left(\sup_{\substack{f_1, f_2 \in \mathcal{F} \\ \|f_1 - f_2\| \leq \delta}} |Z_n(f_1) - Z_n(f_2)| > \eta \right) \leq \eta. \tag{10.2}$$

Then the class $\mathcal{F}$ is P-Donsker.

Proof. Let $\delta > 0$ be fixed. Since the set $\mathcal{F}$ is totally bounded, one can find a finite subset $\mathcal{F}_\delta$ of $\mathcal{F}$ such that for all $f \in \mathcal{F}$, $\exists f_\delta \in \mathcal{F}_\delta$ such that $\|f - f_\delta\| \leq \delta$. Let $k = |\mathcal{F}_\delta|$. Let $h : \mathbb{D} \to \mathbb{R}$ be a Lipschitz function (i.e. $\forall x, y \in \mathbb{D}, |h(x) - h(y)| \leq \lambda \|x - y\|_{\mathbb{D}}$). We define a function $\Pi_\delta : \mathcal{F} \to \mathcal{F}_\delta$ such that $\forall f$, $\Pi_\delta(f)$ is an element of $\mathcal{F}_\delta$ such that $\|f - f_\delta\| \leq \delta$ (in case of multiple choices, one just choose one of them arbitrarily). The process $Z_n \circ \Pi_\delta$ is no more than a random vector of length k. Indeed, if we denote by $f_1, \ldots, f_k$ the elements of $\mathcal{F}_\delta$, then $Z_n \circ \Pi_\delta$ can only take the finitely many possible values $Z_n(f_1, \ldots, Z_n(f_k)$. Then there exists a Lipschitz function $\tilde{h} : \mathbb{R}^k \to \mathbb{R}$ such that

$$h(Z_n \circ \Pi_\delta) = \tilde{h}(Z_n \circ \Pi_\delta).$$

By the classical TCL theorem on vectors, $Z_n \circ \Pi_\delta \xrightarrow{(d)} \mathcal{N}(0, \Sigma)$ where $\Sigma_{f,g} = Pfg - PfPg$ for all $f, g \in \mathcal{F}_\delta$. Now define a centered Gaussian process G such that for all $f, g \in \mathcal{F}$,

$$\operatorname{Cov}(G(f), G(g)) = Pfg - PfPg$$

then $G \circ \Pi_\delta$ is a Gaussian vector which have the same distribution has $\mathcal{N}(0, \Sigma)$. Then $Z_n \circ \Pi_\delta \xrightarrow{(d)} G \circ \Pi_\delta$ and

$$\mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right] = \mathbb{E}\left[\tilde{h}(Z_n \circ \Pi_\delta)\right] \underset{n \to +\infty}{\longrightarrow} \mathbb{E}\left[\tilde{h}(G \circ \Pi_\delta)\right] = \mathbb{E}\left[h(G \circ \Pi_\delta)\right]$$

Now, we show that the condition (10.2) holds also for the process G without the limit in n. Let $\mathcal{G}$ be a finite subset of $\mathcal{F}$. Then by Portmanteau lemma (Lemma 1),

$$\begin{aligned}\mathbb{P}\left(\sup_{\substack{f_1, f_2 \in \mathcal{G} \\ \|f_1 - f_2\| \leq \delta}} |G(f_1) - G(f_2)| > \eta\right) &\leq \liminf \mathbb{P}\left(\sup_{\substack{f_1, f_2 \in \mathcal{G} \\ \|f_1 - f_2\| \leq \delta}} |Z_n(f_1) - Z_n(f_2)| > \eta\right) \\ &\leq \liminf \mathbb{P}\left(\sup_{\substack{f_1, f_2 \in F \\ \|f_1 - f_2\| \leq \delta}} |Z_n(f_1) - Z_n(f_2)| > \eta\right) \\ &\leq \limsup \mathbb{P}\left(\sup_{\substack{f_1, f_2 \in F \\ \|f_1 - f_2\| \leq \delta}} |Z_n(f_1) - Z_n(f_2)| > \eta\right) \leq \eta.\end{aligned}$$

Since the last inequality is true for any finite set $\mathcal{G}$, it is also true for any $\mathcal{G}$ countable. Proceeding as in the proof of Theorem 19, since $\mathcal{F}$ is totally bounded, one can find a countable set $\mathcal{G}$ that is dense inside of $\mathcal{F}$. Hence by defining $\tilde{G}$ the random process on $\mathcal{F}$ defined as $\tilde{G}(f) = G(f)$ for any $f \in \mathcal{G}$ and $\lim G(f_n)$ where $f_n \to f \in \mathcal{F}$ otherwise. This construct a a.s. continuous process $\tilde{G}$ that is a modification of G and such that

$$\mathbb{P}\left(\sup_{\substack{f_1, f_2 \in \mathcal{F} \\ \|f_1 - f_2\| \leq \delta}} |\tilde{G}(f_1) - \tilde{G}(f_2)| > \eta\right) \leq \eta.$$

To conclude the proof by showing that $Z_n \xrightarrow{(d)} \tilde{G}$.

$$\mathbb{E}\left[h(Z_n)\right] - \mathbb{E}\left[h(\tilde{G})\right] = \left(\mathbb{E}\left[h(Z_n)\right] - \mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right]\right) + \left(\mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right] - \mathbb{E}\left[h(G \circ \Pi_\delta)\right]\right) + \left(\mathbb{E}\left[h(\tilde{G} \circ \Pi_\delta)\right] - \mathbb{E}\left[h(\tilde{G})\right]\right)$$

But

$$\begin{aligned}\left|\mathbb{E}\left[h(Z_n)\right] - \mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right]\right| &\leq \lambda\eta + 2\|h\|_\infty \mathbb{P}\left(\|Z_n - Z_n \circ \Pi_\delta\|_{\mathbb{D}} > \eta\right) \\ &\leq \lambda\eta + 2\|h\|_\infty \mathbb{P}\left(\sup_{\substack{f_1, f_2 \in \mathcal{F} \\ \|f_1 - f_2\| \leq \delta}} |Z_n(f_1) - Z_n(f_2)| > \eta\right)\end{aligned}$$

which implies that the term $\left|\mathbb{E}\left[h(Z_n)\right] - \mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right]\right|$ tends to 0 when $n \to \infty$ by choosing η and δ close enough to 0. The exact same argument can be repeated on the last term in the decomposition by using the equicontinuity condition that we showed on $\tilde{G}$. The first part of the proof already showed that $\mathbb{E}\left[h(Z_n \circ \Pi_\delta)\right] - \mathbb{E}\left[h(G \circ \Pi_\delta)\right] \to 0$. This finishes the proof. □

In facts, the condition (10.2) on a totally bounded set $\mathcal{F}$ imposes that the trajectories are a.s. bounded and then belong to $\ell_\infty(\mathcal{F})$. This can be shown by using Borel-Cantelli Lemma and noticing that the process Z_n is trivially bounded on a finite subset of $\mathcal{F}$. Lemma 20 only deals with the case of a random process Z_n with trajectories belonging to $\ell_\infty(\mathcal{F})$. The ideas can be adapted to the other natural metric space described above. For example, if the trajectories of Z_n are elements of $L_p(\mathcal{F}, F)$ where F is a probability measure on $\mathcal{F}$, one have to assume that

$$\limsup \mathbb{P}\left(\int_{\|f_1 - f_2\| \leq \delta} (Z_n(f_1) - Z_n(f_2))^p dF > \eta\right) \leq \eta.$$

10.2 P-Donsker theorems

In this section, we derive a theorem under a condition of a control of the empirical entropy over the class $\mathcal{F}$.

Theorem 26. *We assume that $(\mathcal{F}, \|\cdot\|) \subset L_2(P)$ have a enveloppe $F \in L_2(P)$ and that there exists a non-decreasing function H such that*

1. $\int_0^1 \sqrt{H(\delta)}d\delta < \infty$.

2. $\lim_{A\to\infty} \limsup \mathbb{P}\left(\sup_{\delta>0} \frac{H_2(\delta,\mathcal{F},P_n)}{H(\delta)} > A\right) = 0$.

Then, for all $\eta > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\substack{f_1,f_2\in\mathcal{F}\\ \|f_1-f_2\|\le\delta}} |Z_n(f_1) - Z_n(f_2)| > \eta\right) \le \eta.$$

Proof. We define the events $A_\delta = \{\sup_{\|f_1-f_2\|\le\delta} |Z_n(f_1) - Z_n(f_2)| > \eta\}$ and $E_{n,A} = \{\sup_{\delta>0} \frac{H_2(\delta,\mathcal{F},P_n)}{H(\delta)} > A\}$. Then

$$\begin{aligned}
\mathbb{P}(A_\delta) &= \mathbb{E}\left[\mathbb{1}_{A_\delta}\mathbb{1}_{E_{n,A}}\right] + \mathbb{E}\left[\mathbb{1}_{A_\delta}\mathbb{1}_{E^c_{n,A}}\right] \\
&\le \mathbb{P}(E_{n,A}) + \frac{\mathbb{E}\left[\sup_{\|f_1-f_2\|\le\delta} |Z_n(f_1) - Z_n(f_2)|\mathbb{1}_{E^c_{n,A}}\right]}{\eta} \\
&= \mathbb{P}(E_{n,A}) + \frac{\mathbb{E}\left[\sup_{\|f_1-f_2\|\le\delta} |Z_n(f_1 - f_2)|\mathbb{1}_{E^c_{n,A}}\right]}{\eta} \\
&\le \mathbb{P}(E_{n,A}) + 24\frac{\mathbb{E}\left[\int_0^{\delta_n/2} \sqrt{H_2(\varepsilon,\mathcal{G},P_n)}d\varepsilon\ \mathbb{1}_{E^c_{n,A}}\right]}{\eta}
\end{aligned}$$

where $\mathcal{G} = \{g = f_1 - f_2 : f_1, f_2 \in \mathcal{F} \text{ and } \|g\| \le \delta\}$ and $\delta_n = \sup_{g\in\mathcal{G}} \|g\|_{n,2}$. First of all, a covering of $\mathcal{F}$ by balls of radius ε gives directly a covering of $\mathcal{G}$ by balls of radius 2ε. Indeed, if we denote by $f_1, \dots, f_k$ the centers for a covering of $\mathcal{F}$, then the functions $g_{i,j} = f_i - f_j$ define the centers a covering for $\mathcal{G}$ where the radius has to be multiplied by 2. Then $H_2(2\varepsilon, \mathcal{G}, P_n) \le 2H_2(\varepsilon, \mathcal{F}, P_n)$. We are now ready to apply theorem 18 to the empirical process $(\|f - f'\|^2_{n,2})_{f,f'\in\mathcal{F}}$. For any x, $(f - f')^2(x) \le 2F(x)(f - f')(x)$ so that $\|f - f'\|^2_{n,2} \le 2\|F\|_{n,2}\|f - f'\|_{n,2}$. By the convergence given by the law of large numbers $\|F\|_{n,2} \xrightarrow{a.s.} \|F\|_2$ we have that for n large enough, $\forall f, f' \in \mathcal{F}$,

$$\|f - f'\|^2_{n,2} \le 4\|F\|_2\|f - f'\|_{n,2}.$$

Let

$$\mathcal{H} = \{(f - f')^2 : f, f' \in \mathcal{F}\}.$$

So we have that

$$H_2(\varepsilon, \mathcal{H}, P_n) \le H_2(\frac{\varepsilon}{4\|F\|_2}, \mathcal{G}, P_n) \le 2H_2(\frac{\varepsilon}{8\|F\|_2}, \mathcal{F}, P_n).$$

But the enveloppe of the class $\mathcal{H}$ is $x \mapsto \sup_{f_1,f_2}(f_1 - f_2)^2(x) \le 4F^2(x)$ is in $L_1(P)$ and then

$$\sup_{f_1,f_2} \left|\|f_1 - f_2\|^2_{2,n} - \|f_1 - f_2\|^2_2\right| \xrightarrow{a.s.} 0$$

so that for n large enough, for every $f_1, f_2 \in \mathcal{F}$,

$$\frac{1}{2}\|f_1 - f_2\|_{2,n} \le \|f_1 - f_2\|_2 \le 2\|f_1 - f_2\|_{2,n}$$

and then for n large enough, $H_2(2\varepsilon, \mathcal{F}, P) \le H_2(\varepsilon, \mathcal{F}, P_n)$ which implies that $\mathcal{F}$ is totally bounded. The same result also implies that $\delta_n \le 2\delta$, for all n large enough. Finally,

$$\mathbb{P}(A_\delta) \le \mathbb{P}(E_{n,A}) + 48\sqrt{2}\frac{\mathbb{E}\left[\int_0^\delta \sqrt{H_2(\varepsilon,\mathcal{F},P_n)}d\varepsilon\ \mathbb{1}_{E^c_{n,A}}\right]}{\eta} \le \mathbb{P}(E_{n,A}) + 48\sqrt{2}\frac{\mathbb{E}\left[\int_0^\delta \sqrt{H(\varepsilon)}d\varepsilon\ \right]}{\eta}$$

We conclude the proof by taking $n \to \infty, A \to \infty$ and $\delta \to 0$. □

Chapter 11

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].

11.1 Notations and definitions

11.1.1 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. $0 \leq x_i < 1$, $i = 1, \dots, m$). We denote by $k = (k_1, \dots, k_m)$ 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

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

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

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

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

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

11.1.2 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 $\Lambda = \{\Delta_1, \dots, \Delta_{|\Lambda|}\}$. 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

$$\begin{array}{rl} \textbf{Minimize} & \Lambda \mapsto G_a(J, \Lambda) \\ \textbf{where} & |\Lambda| \leq n. \end{array} \tag{11.1}$$

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. Assuming the partition Λ_i already constructed, we slice the cubes Δ for which $g_a(J, \Delta)$ is such that

$$g_a(J, \Delta) \geq 2^{-ma} G_a(J, \Lambda_i) \tag{11.2}$$

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

11.1.3 Two elementary lemmas

Lemma 21. *Suppose that a cube Δ in Q^m is divided into cubes Δ_j for $j = 1, \ldots, 2^m$. 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 22. *Let $s \in \mathbb{N}$ and let $x_j > 0$, $y_j > 0$ $(j = 1, \ldots, s)$ be numbers such that*

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

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

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

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. □

11.2 Bounding the entropy of $\mathcal{J}$

11.2.1 The decreasing behavior of G_a

In this section, we investigate the effect of the Slicing strategy over the decreasing behavior of the functional G_a. Since the slicing is performed on all the cubes of 'large' weight for G_a, we expect to see a decreasing effect on the G_a along the successive refinement of the partitions. A precise statement is as follows. For this purpose, we define the quantities

$$\forall i \geq 0,\ \ n_i = |\Lambda_i|$$

on which can directly see that $n_0 = 1$ and

$$n_{i+1} \leq 2^m n_i.$$

We define

$$S_i = \{j \in \{1, \ldots, n_i\} : g_a(J, \Delta_j) \geq 2^{-ma}\delta_i\} \quad \text{and} \quad s_i = |S_i|$$

for the cubes that are sliced (or equivalently that satisfy (11.2)) at step i to obtain the partition Λ_{i+1}.

Lemma 23. *We have the relation for any $i \geq 1$,*

$$n_i \leq 2^m \sum_{j=0}^{i-1} s_j.$$

Proof. When a cube is sliced, then one cube disappear and 2^m cubes are created. Since we do it for s_i cubes, we get $n_{i+1} - n_i = (2^m - 1)s_i$. Summing this last equality gives the result. □

Theorem 27. *Assume that the function J is semiadditive from below. Then there exists a constant $C = C(a, m)$ (that do not depend on J) such that for all $i \geq 1$,*

$$G_a(J, \Lambda_i) \leq C n_i^{-(a+1)}. \tag{11.3}$$

In particular, for $C_1 = C_1(a, m) = 2^{m(a+1)}C(a, m)$ then for all $n \in \mathbb{N}^*$, there exists a partition Λ of the cube Q^m of size $|\Lambda| \leq n$ and such that

$$G_a(J, \Lambda) \leq C_1 n^{-(a+1)} J(Q^m).$$

Proof. The second part on the theorem is a direct consequence of (11.3) so we only have to prove (11.3). For simplicity of the notations, we note $\delta_i = G_a(J, \Lambda_i)$ in the following. The partition strategy and the quantitative result in Lemma 21 show that $\forall i \geq 0$,

$$\delta_{i+1} \leq 2^{-ma}\delta_i. \tag{11.4}$$

But using Lemma 22 for the quantities $x_j = J(\Delta_j)$, $y_j = |\Delta_j|$ over the class of cubes corresponding to S_i and $b = 2^{-ma}\delta_i$, we get

$$\delta_i \leq 2^{ma} s_i^{-(a+1)} \tag{11.5}$$

Equations (11.4) and (11.5) together show that for any $k \geq i$,

$$\delta_k \leq 2^{-(k-i-1)ma} s_i^{-(a+1)}$$

or again

$$s_i \leq \delta_k^{-(a+1)^{-1}} \times 2^{-(k-i-1)ma(a+1)^{-1}}.$$

Summing those relations over $i \in \{0, \ldots, k-1\}$ and denoting $q = 2^{-ma(a+1)^{-1}}$, we have

$$n_k \leq 2^m \sum_{i=0}^{k-1} s_i \leq 2^m \delta_k^{-(a+1)^{-1}} \sum_{i=0}^{k-1} q^i \leq 2^m \delta_k^{-(a+1)^{-1}} \sum_{i=0}^{\infty} q^i = 2^m \frac{\delta_k^{-(a+1)^{-1}}}{1-q}$$

Then $\delta_k \leq (2^m/(1-q))^{-(a+1)} \times n_k^{-(a+1)}$ which gives the result for the constant $C = (2^m/(1 - 2^{-ma(a+1)^{-1}}))^{-(a+1)}$. □

11.2.2 The one dimensional case

In this section, we see that Theorem 27 to the one dimensional case. The change is that a is allowed to take the value 0 in this case. We denote by $I = [0,1)$ and for any $x, y \in [0,1)$ such that $x < y$ and denote the semiadditive function J taken on the half open intervals by $J[x,y)$. By the semiadditive property, we see that the function $\phi(t) \mapsto J[t,y)$ is non-increasing on (x,y) and bounded. We define

$$\widetilde{J}[x,y) = \lim_{t\to x^+} \phi(t).$$

This definition implies that $\widetilde{J}[x,y) \leq J[x,y)$.

Theorem 28. *Assume that the function J is non-negative, semiadditive from below and continuous on the left which is*

$$\lim_{t\to y^-} J[x,t) = J[x,y).$$

Then, for any $a \geq 0$ and any $n \in \mathbb{N}$, there exists a partition Λ of the interval with $|\Lambda| \leq n$ and

$$G_a(\widetilde{J}, \Lambda) \leq n^{-(a+1)} J[0,1).$$

Chapter 12

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.

12.1 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 $\mathbb{E}\left[\gamma_n(s)\right]$. The purpose of M-estimation is exactly to define a random point that we hope to be close.

Definition 13. *We define the following notions.*

1. *Let s be the* ***target*** *parameter defined as*
$$s \in \operatorname{argmin}_{t \in \mathcal{S}} \mathbb{E}\left[\gamma_n(t)\right].$$
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*
$$R(t) = \mathbb{E}\left[\gamma_n(t)\right]$$
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

$$P f = \mathbb{E}\left[f(X)\right]$$
$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

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

12.2 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 26. *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).

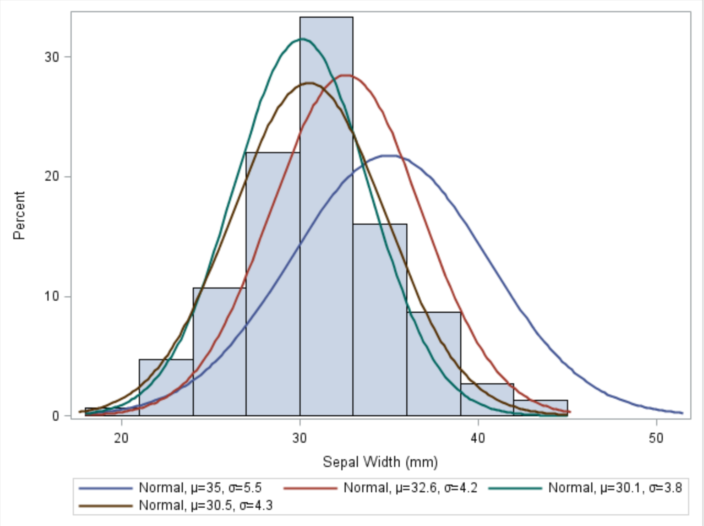


Figure 12.1: An example of MLE done by hands.

12.3 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 14. *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 [15], [16] or [10].

12.3.1 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 15 (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 27. *Prove the equivalences in Definition 15.*

The equivalences in Definition 15 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 24. *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 13

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 [12].

13.1 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 16. *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,*

   $$m_* \in \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\gamma_n(\hat{s}_m)] \tag{13.1}$$

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 (13.1)) 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 (13.1) 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 $\hat{s}_m$,

$$\begin{aligned} \mathbb{E}_\gamma[\gamma_n(\hat{s}_m)] &\quad \text{is a random variable} \\ \mathbb{E}[\gamma_n(\hat{s}_m)] &\quad \text{is a deterministic number} \end{aligned}$$

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

$$\hat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \gamma_n(\hat{s}_m) \tag{13.2}$$

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

$$\mathbb{E}[\overline{\gamma}_n(\hat{s}_m)] \leq 0. \tag{13.3}$$

This can be interpreted as the fact that the minimization in (13.2) 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.

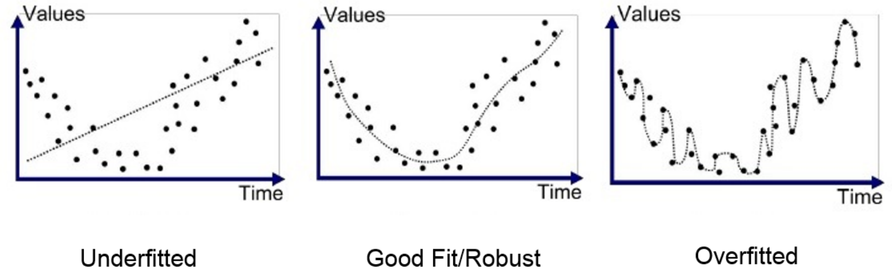


Figure 13.1: A typical problem of underspecified (left) vs adapted (center) vs over-fitting (right)

13.1.1 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 17. *A penalization on the class of models* $(S_m)_{m\in\mathcal{M}}$ *is a function* $\mathrm{pen}:\mathcal{M}\to\mathbb{R}_+$. *We allow* $\mathrm{pen}(m)$ *to be a random variable depending on the data* $X_1,\dots,X_n$.

The new estimator is then defined as a minimPseizer of

$$\hat{m}\in\mathrm{argmin}_{m\in\mathcal{M}}\{\gamma_n(\hat{s}_m)+\mathrm{pen}(m)\}. \tag{13.4}$$

For clarity in the notations, from now on, we denote by $\tilde{s}$ the **selected** estimator using (13.4) estimator $\hat{s}_{\hat{m}}$.

Ideal penalizations

We define the ideal penalizations

$$\mathrm{pen}^{id}_{det}(m)=\mathbb{E}\left[\gamma_n(\hat{s}_m)\right]-\gamma_n(\hat{s}_m) \tag{13.5}$$
$$\mathrm{pen}^{\mathrm{id}}_{ran}(m)=\mathbb{E}_\gamma\left[\gamma_n(\hat{s}_m)\right]-\gamma_n(\hat{s}_m)=-\overline{\gamma}_n(\hat{s}_m). \tag{13.6}$$

In practice, $\mathrm{pen}^{\mathrm{id}}$ 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 $\mathrm{pen}=\mathrm{pen}^{id}_{det}$, then it is clear that $\hat{m}=m_*$ and this choice would achieve the prefect estimator $\hat{s}_{m_*}$.

13.1.2 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

$$\ell(\tilde{s},s)\le C\inf_{m\in\mathcal{M}}\ell(\hat{s}_m,s)+\mathrm{Dev}=C\ell(\hat{s}_{m_*},s)+\mathrm{Dev} \tag{13.7}$$

for C a positive constant. A result as (13.7) 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 (13.1) 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

$$\ell(\tilde{s},s)\le C\inf_{m\in\mathcal{M}}\left(\ell(s_m,s)+\mathrm{pen}(m)\right)+\mathrm{Dev}' \tag{13.8}$$

where the infimum describes the best possible projection on S_m weighed by the penalization term.

A generic calculation We have, from (13.6), the following calculations

$$\begin{aligned}\ell(\tilde{s}, s) &= \mathbf{E}\left[\gamma_n(\tilde{s})\right] - \mathbf{E}\left[\gamma_n(s)\right] \\ &= \gamma_n(\tilde{s}) + \mathrm{pen}^{\mathrm{id}}(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \\ &= \gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \\ &\leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) - \mathbf{E}\left[\gamma_n(s)\right] \end{aligned} \tag{13.9}$$

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

$$\ell(\tilde{s}, s) \leq \ell(\hat{s}_m, s) + (\mathrm{pen} - \mathrm{pen}^{id})(m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) \tag{13.10}$$

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 (13.9) and using

$$\gamma_n(\hat{s}_m) \leq \gamma_n(s_m),$$

the bound on $\ell(\tilde{s}, s)$ becomes

$$\ell(\tilde{s}, s) \leq \ell(s_m, s) + \mathrm{pen}(m) + \overline{\gamma}_n(s_m) + (\mathrm{pen}^{\mathrm{id}} - \mathrm{pen})(\hat{m}) \tag{13.11}$$

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 (13.7). For example, if one can control uniformly the deviation between $\mathrm{pen}_{\mathrm{id}}$ and pen with high probability,

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \mathrm{pen}(m) \leq \mathrm{pen}_{\mathrm{id}}(m) + C \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s) \tag{13.12}$$

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,

$$|\,\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)| \leq \varepsilon \inf_{m \in \mathcal{M}} \ell(\hat{s}_m, s) \tag{13.13}$$

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 $n \to \infty$.

Chapter 14

Extra definitions

14.0.1 Sumable familly

Definition 18. *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 30. *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.

14.0.2 Processes

Definition 19. *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 15

Functional Analysis

15.1 Lemmas

We give here the proofs of some technical results.

Lemma 25. *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 26. *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 27. *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 28. *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 $\forall x \in I$, $c(x) \geq D(x-v) + c(v)$.

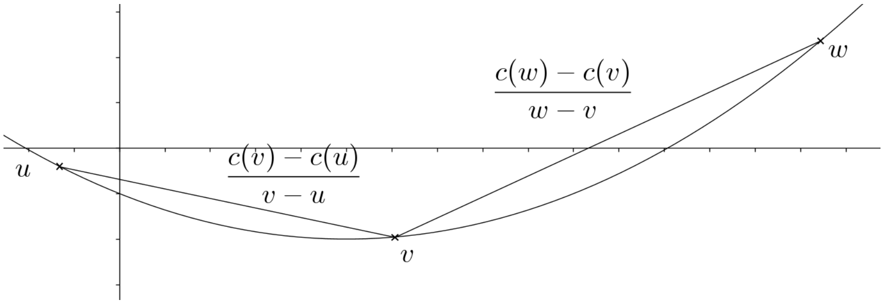


Figure 15.1: Inequality (15.1)

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

$$\Delta_{u,v} \leq \Delta_{u,w} \leq \Delta_{v,w} \quad \text{where} \quad \Delta_{x,y} = \frac{c(y)-c(x)}{y-x}. \tag{15.1}$$

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

$$|C(v) - C(v_0)| = |\Delta_{v_0,v}||v-v_0| \leq \max\{|\Delta_{v_0,w}|, |\Delta_{u,v_0}|\}|v-v_0| \xrightarrow[v\to v_0]{} 0$$

and a) is proved. From (15.1), 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. □

15.2 Basic facts on integrable functions

Proposition 31. *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 9. *Let $f \geq 0$ be an integrable function, then*

$$\int f(x)\mathbb{1}_{|f(x)|>t} dx \xrightarrow[t\to\infty]{} 0.$$

15.3 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 32. *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 33. *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*

$$f \star \phi_n \underset{n \to \infty}{\longrightarrow} f \quad (\textit{in } L_p(\mathbb{R}^k)).$$

To prove that fact, we begin with a stronger case.

Lemma 29. *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 33.

Proof of Proposition 33. 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

$$\|(f \star \phi_n) - f\|_p^p \leq \int_{\mathbb{R}^k} \|f_y - f\|_p^p \, \phi_n(y)dy, \tag{15.2}$$

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 (15.2) as $g \star \phi_n(0)$ we get, by Lemma 29, that it converges to 0 as $n \to +\infty$. □

15.4 Distribution functions and simple functions

Definition 20. *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 21. *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 29. *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 [11, Theorem 7.18] □

We have the useful lemma:

Lemma 30. *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*

$$\int_{\mathcal{X}} (\phi \circ f)d\mu = \int_0^{+\infty} \mu\{f > t\}\phi'(t)dt \tag{15.3}$$

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 (15.3) for any measurable function. □

A special case of Lemma 30 is the following result.

Corollary 10. *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

$$\int_0^{+\infty} \mathbb{P}(X \geq t)\, dt \tag{15.4}$$

since integration on the open $(0, +\infty)$ or on $[0, +\infty)$ are equivalent for the Lebesgue measure dt.

Proof. Apply Lemma 30 for f, ϕ both equal to the identity function. □

15.5 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

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

15.5.1 Dominated convergence

Theorem 30 (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,

$$f_n \xrightarrow{\mathbb{L}_1} f \ \textit{in}\ L_1(\mathcal{X}, \mu),$$

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 11 (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 30 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),

$$\limsup \mathbb{1}_{Z_n \geq \varepsilon} = \lim_{n\to\infty} \sup_{k \geq n} \mathbb{1}_{Z_k \geq \varepsilon} = \lim_{n\to\infty} \mathbb{1}_{\sup_{k\geq n}(Z_k) \geq \varepsilon} = \mathbb{1}_{\lim\limits_{n\to\infty} \sup_{k\geq n}(Z_k) \geq \varepsilon}$$

where the last equality is clear since the sequence $(\sup_{k\geq n}(Z_k))_n$ is monotone. □

Lemma 31 (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 12. *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 32 (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,

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

□

15.5.2 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 33 (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 34 (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 $(g - f_n)_n$. □

15.6 The Monotone convergence theorem

15.6.1 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 35 (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 $\mu(G_k) = \mu(G) + \mu(G_k \backslash G)$. □

15.6.2 Technical lemmas

Doubly monotone convergence

Lemma 36 (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 22. *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 $\alpha_p(x) \to \alpha_p(x_0)$).

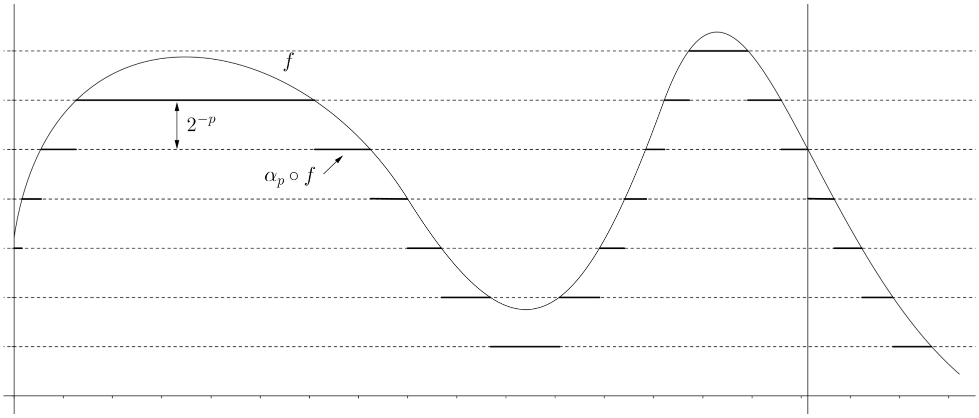


Figure 15.2: An example of the staircase transformation

Proposition 34. *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 37. *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 35, $\mu(A_n) \uparrow \mu(A)$. 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

$$\overline{f}_k = \max\{g_k, h_1, \dots, h_k\}.$$

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 36 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 31 ((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 37, 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 36 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 16

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 38 (Borel-Cantelli)**.** *For a sequence of events $(E_n)_n$ such that $\sum_{n\geq 0} \mathbb{P}(E_n) < +\infty$. Then*

$$\mathbb{P}(\limsup E_n) = \mathbb{P}(E_n \ i.o.) = 0$$

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

$$\mathbb{P}(G) \underset{\text{Lemma 35}}{\leq} \mathbb{P}(G_m) \leq \sum_{n\geq m} \mathbb{P}(E_n).$$

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 39 (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*

$$\mathbb{P}(\limsup E_n) = \mathbb{P}(E_n \ i.o.) = 1$$

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 40 (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 28, 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. □

16.0.1 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 41. *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 41 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 42. *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 41. Let $\varepsilon_k = 2^{-k}$. The convergence in probability implies that $\mathbb{P}\left(|X_n - X| > \varepsilon_k\right) \underset{n\to\infty}{\longrightarrow} 0$ 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 41,

$$\sum_{k=0}^{+\infty} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right) \leq \underbrace{\sum_{k=0}^{k_0-1} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right)}_{<+\infty} + \sum_{k=k_0}^{+\infty} \underbrace{\mathbb{P}\left(|X_{n_k} - X| > \varepsilon_k\right)}_{\text{summable}} < +\infty$$

and then $X_{n_k} - X \xrightarrow{a.s.} 0$. □

16.0.2 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 43 (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) $\mathbb{P}(X_n = Y_n \text{ eventually}) = 1$. *[PRECISE THIS]*

ii) $\|\sum_{n\geq 1} n^{-2}\, Var(Y_n)\| < \infty$.

Proof. For proving i), we use Borel-Cantelli's lemma (Lemma 38) 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 44. *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 43, 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 43 allows us to use Lemma 41 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 43 will be replaced by $\mathbb{P}\left(\forall t\in\mathcal{T},X_{i,t}=Y_{i,t}\text{ eventually}\right)=1$.

Exercice 28. *Show the equivalence of Remark 4.*

16.1 Stein equation for Gaussian vectors

Stein equation is one of the numerous characterization of the Gaussian law. We first derive the 1 dimensional case and extend the sufficient condition in the case of Gaussian vectors. We recall that the notation $\mathcal{C}_b^1$ holds for the set of functions that are piecewise differentiable and of bounded derivative.

Proposition 35. *Let X be a real random variable such that $\mathbb{E}\left[X\right]=0$ and $Var(X)=\sigma^2$. Then it holds that*

$$\mathbb{E}\left[XF(X)\right]=\sigma\mathbb{E}\left[F'(X)\right],\ \forall F\in\mathcal{C}_b^1(\mathbb{R})\qquad\Leftrightarrow\qquad X\text{ is Gaussian}\tag{16.1}$$

Proof. Without loss of generality, we only prove the Proposition for standard Gaussian variables as it is always possible to renormalize a centered Gaussian variable to a standard one. Let Z be a standard Gaussian variable, then for any function F in $\mathcal{C}_b^1(\mathbb{R})$,

$$\begin{aligned}
\mathbb{E}\left[F'(Z)\right] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} F'(z)e^{-z^2/2}dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} F'(z)\int_{-\infty}^{z} -xe^{-x^2/2}dxdz + \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} F'(z)\int_{z}^{+\infty} xe^{-x^2/2}dxdz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}\left(\int_{x}^{0} F'(z)dz\right)(-x)e^{-x^2/2}dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty}\left(\int_{0}^{x} F'(z)dz\right)xe^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(F(x)-F(0))xe^{-x^2/2}dx = \mathbb{E}\left[ZF(Z)\right].
\end{aligned}$$

and this shows the necessary condition. For the sufficient condition, let F_z be a solution of the differential equation

$$y' - xy = \mathbb{1}_{(-\infty,z]}(x) - \Phi(z)$$

where Φ is the cumulative distribution function for the standard Gaussian variable. This simple differential equation have solutions that are in $\mathcal{C}_b^1$ and can even be expressed explicitly in terms of the function Φ. Since $\mathbb{E}\left[XF(X)\right] = \mathbb{E}\left[F'(X)\right]$ for any function in $\mathcal{C}_b^1$, one deduces that

$$0 = \mathbb{E}\left[F_z'(X) - XF_z(X)\right] = \mathbb{P}\left(X \le z\right) - \Phi(z)$$

which concludes the proof. □

It is possible to generalize the sufficient condition of Proposition 35 to a multidimensional context.

Proposition 36. *Let $F : \mathbb{R}^d \to \mathbb{R}$ a function in $\mathcal{C}_b^1$ and let $X = (X_1, \ldots, X_d)$ be a centered Gaussian vector. Then, for any $1 \le i \le d$*

$$\mathbb{E}\left[X_iF(X)\right] = \sum_{j=1}^{d}\mathbb{E}\left[X_iX_j\right]\mathbb{E}\left[\partial_jF(X)\right].$$

Proof. This is easily proved using Proposition 35 and a conditioning on the variables X_k for $k \neq i$. □

Chapter 17

Carathéodory theorem

17.1 Measure set theory

17.1.1 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.
- $F_1, F_2, \dots, F_n, \dots \in \Sigma$ 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 $G_1 = F_1, G_2 = F_2 \backslash F_1, \dots, G_n = F_n \backslash \bigcup_{i=1}^{n-1} F_i, \dots$ 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

$$\forall K \in \Sigma_0,\ \lambda(L \cap K) + \lambda(L^c \cap K) = \lambda(K). \tag{17.1}$$

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 $\mathcal{C}$.

17.1.2 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,

$$\mu_0(\bigcup_{n\in\mathbb{N}} F_n) \leq \sum_{n\in\mathbb{N}} \mu_0(F_n).$$

17.1.3 Extension theorems

Proposition 37 (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

$$\begin{aligned} \lambda(L \cap K) + \lambda(L_1^c \cap L_2 \cap K) &= \lambda(L_2 \cap K) && \text{(with } L_1 \text{ and } L_2 \cap K\text{)} \\ \lambda(L_2 \cap K) + \lambda(L_2^c \cap K) &= \lambda(K) && \text{(with } L_2 \text{ and } K\text{)} \\ \lambda(L^c \cap K) &= \lambda(L_2 \cap L_1^c \cap K) + \lambda(L_2^c \cap K) && \text{(with } L_2 \text{ and } L^c \cap K\text{)} \end{aligned}$$

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 (17.1) 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 45 (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 37, 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 $L_1, \dots, L_n, \dots$ 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 37 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

$$\lambda(L) = \sum_{k\geq 1} \lambda(L_k).$$

□

17.1.4 Carathéodory theorem

The following theorem is an angular stone to construct all the measures that are commonly used in probabilistic theory.

Theorem 32. *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 32. 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

$$G_n \subseteq \bigcup_{k\geq 1} F_{n,k} \quad \text{and} \quad \sum_{k\geq 1} \mu_0(F_{n,k}) < \lambda(G_n) + \varepsilon 2^{-n}.$$

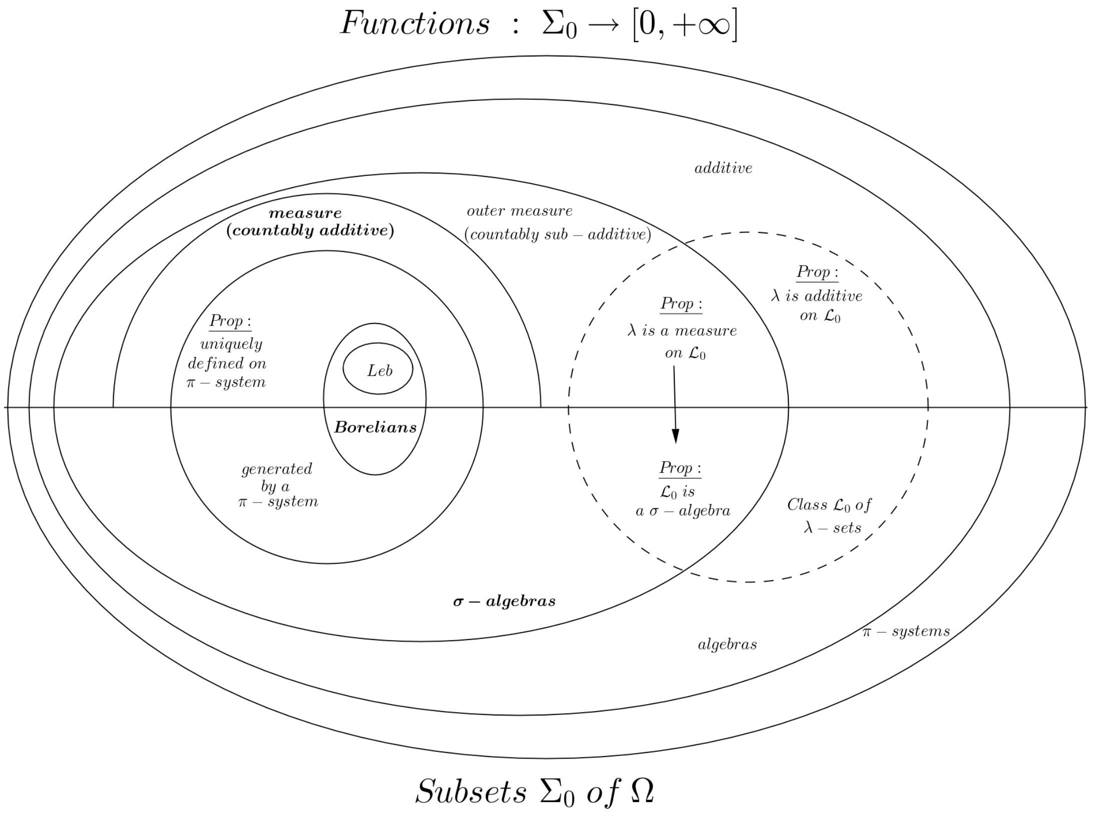


Figure 17.1: A sum up of the classes of importance in measure theory represented as inclusion of sets for sub-classes. On the bottom side, the definitions of different types of classes correspond to definitions for non-negative valued function on the top. The inclusions represents sub-classes and bold 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 45, 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 33 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

$$\lambda(G) = \inf \sum_{n\geq 1} \mu_0(F_n) \quad (\forall G \in \mathcal{P}(\Omega)) \tag{17.2}$$

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.

17.1.5 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 38. *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 46 (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 33 (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 35. 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 35. The conclusion also fails to hold if μ_1 and μ_2 are only assumed to be finitely additive.

17.1.6 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 (17.2) equals μ_0 on the algebra. We follow the second option here. The interested reader may find the other option in [18, 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 32, 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 $\lambda(\{a\}) = 0$.

17.2 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 $X = \mathrm{id}$ 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{Leb})$). This new question is tackled by Skorohod construction.

17.2.1 Real valued random variables

Chapter 18

Szemeredi Regularity Lemma

18.1 A basic lemma

A refined version of Cauchy-Schwarz inequality One can use regular Cauchy-Schwarz inequality to obtain the following refined result.

Lemma 47. *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 13. *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 47 with $a_i = 1/n$, $\mu = \delta/m$ and $b_i = x_i$. □

18.2 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 23. *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 24. *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 [13] 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 48. *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 34 (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 34. 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 48, 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 $N(\varepsilon, t) = f(M(\varepsilon, t)) + 1$. □

Chapter 19

Tareas

19.1 Tarea 1

Esa tarea está dividida en problemas independientes. Fecha limite de entrega : **09/03/2020**

Problema 1 (Alrededor de funciones caracteristicas) Sea Z una variable uniforme sobre $[-1, 1]$.

1. Calcular la función caracteristica de Z.
2. Mostrar que no se puede encontrar variables i.i.d. X, Y tal que $X - Y \sim Z$.

Sea $f : t \mapsto ae^{b(|t|+c)^2}$.

3. Mostrar que f es una función caracteristica por ciertas constantes a, b, c. Describir la distribución corespondiente.
4. Mostrar que $t \mapsto e^{-|t|^\alpha}$ por $\alpha > 2$ no puede ser una función carateristica.

Problema 2 (Condiciones de Lindeberg-Feller) Sean $X_i \sim U[-a_i, a_i]$ variables uniformes independientes con $\forall i, a_i < a < \infty$.

1. Mostrar que las condiciones de Lindeberg-Feller se cumplen por la sucesión $(X_i)_i$ si y solo si $\sum_i a_i^2 = \infty$

Sean $X_i \sim \text{Exp}(\lambda_i)$ y supongamos que $(\max_{1\leq i\leq n} \lambda_i^2)/\sum_{i=1}^n \lambda_i^2 \to 0$.

2. Mostrar que bajo la buena standardización (de media y varianza), la suma $\sum_i X_i$ converge a $\mathcal{N}(0, 1)$.

Problema 3 (Aplicación de Slutsky)

1. Sean X_n y Y_m variables aleatorias independientes de Poisson de parámetros n y m. Que distribución limite tiene $\frac{X_n - Y_m - (n-m)}{\sqrt{X_n + Y_m}}$ cuando $n, m \to \infty$?

Problema 4 (Uniforme integrabilidad) Supongamos dadas unas variables reales positivas $X_1, \ldots, X_n$ i.i.d. Denotamos $X_{(1)}, \ldots, X_{(n)}$ las estadísticas de orden.

1. Mostrar que si $\mathbb{E}\left[X_1^k\right] < \infty$, se cumple

$$\mathbb{E}\left[X_{(r)}^k\right] \leq \frac{n!}{(r-1)!(n-r)!}\mathbb{E}\left[X_1^k\right].$$

2. Mostrar que si $\mathbb{E}\left[X_1^2\right] < \infty$, la sucesión $(n^{-1}X_{(n)})_n$ es uniformemente integrable.

Sea $(X_n)_n$ una sucesión de variables reales. Sea $f : \mathbb{R}_+ \to \mathbb{R}_+$ una función no decreciente tal que $\frac{f(x)}{x} \underset{n\to\infty}{\longrightarrow} +\infty$. Supongamos que $\mathbb{E}\left[\sup_n f(|X_n|)\right] < \infty$.

3. Mostrar que $(X_n)_n$ es una sucesión uniformemente integrable.

19.2 Tarea 2

In this tarea se busca comparar la técnica de 'regular chaining' con la de 'generic chaining'. La fecha de entrega de la tarea es el **06/04/2020**.

En lo que sigue, T es un espacio métrico y llamamos d la distancia asociada. Digamos que una sucesión $(\mathcal{A}_n)_n$ de particiones crecientes (i.e. $\mathcal{A}_n \subset \mathcal{A}_{n+1}$) de T es **admisible** si $|\mathcal{A}_n| \leq 2^{2^n} =: N_n$ para cada $n \geq 1$ y $|\mathcal{A}_0| = 1$. Por un elemento $t \in T$, se denota $A(t)$ el único elemento de $\mathcal{A}_n$ que contiene t. Usaremos la notación $\Delta(A)$ para designar el diámetro de $A \in \mathcal{A}_n$.

1. Sea $(\mathcal{A}_n)$ una sucesión admisible y $\mathcal{B}_n = \mathcal{A}_{n-1} \times \mathcal{A}_n$ si $n \geq 1$ y $\mathcal{B}_{-1} = \mathcal{B}_0 = \{T\} \times \{T\}$. Mostrar que $(\mathcal{B}_{n-1})_{n\geq 0}$ es admisible por el espacio $T \times T$.

2. Mostrar que si dos sucesiones $\mathcal{B}_n$ y $\mathcal{C}_n$ son admisibles entonces la sucesión $\mathcal{A}_n$ de las particiones dondes los elementos son de la forma $B \cap C$ con $B \in \mathcal{B}_{n-1}$ y $C \in \mathcal{C}_{n-1}$ y tal que $\mathcal{A}_0 = \{T\}$ es admisible.

3. Dado una sucesión admisible $(\mathcal{A}_n)_n$, decir como construir mapeos $\Pi_n : T \to A_n$ tal que $\forall t \in T$
$$d(\Pi_n(t), \Pi_{n+1}(t)) \leq \Delta(A_n(t)).$$

4. Sea $(X_t)_{t\in T}$ un proceso de incrementos sub-Gaussianos y tal que $\forall t \in T$, $\mathbb{E}[X_t] = 0$. Mostrar que $\forall n \geq 0$ y $\forall u > \sqrt{2\log 2}$,
$$\mathbb{P}\left(\sup_{t\in T} \frac{X_{\Pi_{n+1}(t)} - X_{\Pi_n(t)}}{\Delta(A_n(t))} \geq u2^{n/2}\right) \leq N_{n+2}\exp(-u^2 2^n) \leq \exp(-u^2 2^{n-1}).$$
Mostrar que $\sum_{n\geq 0} \exp(-u^2 2^{n-1}) \leq \sum_{n\geq 1} \exp(-\frac{u^2}{2}n) \leq 2\exp(-\frac{u^2}{2})$.

5. Usando 4., mostrar que
$$\mathbb{P}\left(\forall t \in T,\ X_t < u\sum_{n\geq 0} 2^{n/2}\Delta(A_n(t))\right) \geq 1 - 2\exp(-u^2/2)$$
y deducir que existe una constante $L > 0$ universal tal que
$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq L \inf_{\mathcal{A}_n \text{ admisible}} \sup_{t\in T} \sum_{n\geq 0} 2^{n/2}\Delta(A_n(t)).$$
Esa cota se llama cota de **generic chaining**.

6. Para cada n, definimos $e_n = \inf_{\mathcal{A}} \sup_t \Delta(A_n(t))$. Mostrar que $e_n = 2\inf\{\varepsilon : \mathcal{N}(\varepsilon, T, d) \leq N_n\}$. Deducir que existe una constante universal C tal que
$$\inf_{\mathcal{A}_n \text{ admisible}} \sup_{t\in T} \sum_{n\geq 0} 2^{n/2}\Delta(A_n(t)) \leq \sum_{n\geq 0} 2^{n/2} e_n \leq C\int_0^\infty \sqrt{\log \mathcal{N}(\varepsilon, T, d)}d\varepsilon.$$
La cota de **generic chaining** es mejor que la cota de **Dudley**.

7. Sea $(a_i)_{i\geq 1}$ una sucesión t.q. $a_i > 0$, definimos el elipsoide
$$\mathcal{E} = \left\{t \in \ell^2 : t_i > 0 \text{ y } \sum_i \frac{t_i^2}{a_i^2} = 1\right\}.$$
Sea $(g_i)_{i\geq 1}$ una sucesión i.i.d. de variables Gaussianas estandares. Consideramos el proceso $X_t = \sum_{t\in\mathcal{E}} t_i g_i$. Mostrar que
$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \left(\sum_{i\geq 1} a_i^2\right)^{1/2} \leq \left(\sum_{n\geq 0} 2^n a_{2^n}^2\right)^{1/2}.$$

8. En esa pregunta queremos probar que el orden de $\sum_{n\geq 0} 2^{n/2} e_n$ is of order larger than $(\sum_{n\geq 0} 2^n a_{2^n}^2)^{1/2}$. Sea
$$\mathcal{E}_n = \left\{t \in \mathbb{R}^{2^n} : t_i > 0 \text{ y } \sum_i \frac{t_i^2}{a_i^2} = 1\right\}.$$

(a) Mostrar que $e_n(\mathcal{E}_n) \leq e_n(\mathcal{E})$.

(b) Sea B la bola unitaria Euclidiana de $\mathbb{R}^{2^n}$, sea $T \subset \mathcal{E}_n$ un conjunto finito t.q $|T| \leq N_n$ y sea $\varepsilon > 0$. Mostrar que

$$\mathrm{Vol}\left(\cup_{t\in T}(\varepsilon B + t)\right) \leq (2\varepsilon)^{2^n}\,\mathrm{Vol}(B).$$

(c) Mostrar que $\mathrm{Vol}(\mathcal{E}_n) \geq a_{2^n}^{2^n}\,\mathrm{Vol}(B)$. Deducir que $\mathcal{E}_n \subset \cup_{t\in T}(\varepsilon B + t) \implies 2\varepsilon \geq a_{2^n}$.

(d) Finalmente, probar que $e_n(\mathcal{E}) \geq a_{2^n}/2$ y concluir.

En generalidad completa, la cantidad $\inf_{\mathcal{A}_n \text{ admisible}} \sup_{t\in T} \sum_{n\geq 0} 2^{n/2}\Delta(A_n(t))$ siempre tiene el orden de magnitud correcto aun que la cota de Dudley es demasiado conservativa.

19.3 Tarea 3

Digamos que una función $f : \mathcal{X}^n \to [0,\infty)$ tiene la propiedad de ser *acotada por si misma* si para cada i existe una función $f_i : \mathcal{X}^n \to \mathbb{R}$ tal que

$$0 \leq f(x_1,\dots,x_n) - f_i(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) \leq 1$$

y

$$\sum_{i=1}^n f(x_1,\dots,x_n) - f_i(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) \leq f(x_1,\dots,x_n).$$

Notaciones :

1. $\mathrm{Ent}(X) = \mathbb{E}\left[X \log X\right] - \mathbb{E}X \log(\mathbb{E}X)$
2. $X^{(i)} = (X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$ y $\mathbb{E}^{(i)}[.] = \mathbb{E}[.|X^{(i)}]$.
3. $\mathrm{Ent}^{(i)}(X) = \mathbb{E}^{(i)}[X \log X] - \mathbb{E}^{(i)}X \log(\mathbb{E}^{(i)}X)$.
4. $\phi(u) = e^u - 1 - u$.
5. $\psi_{Z-\mathbb{E}Z}(\lambda) = \log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]$

Se **admite** la desigualdad de sub-aditividad de entropías : $\mathrm{Ent}(Z) \leq \mathbb{E}\sum_{i=1}^n \mathrm{Ent}^{(i)}(Z)$.

P.1 Sea $I \subset \mathbb{R}$ un intervalo abierto y $f : I \to \mathbb{R}$ una función convexa y derivable. Sea X una variable tal que $X \in I$. Mostrar que

$$\mathbb{E}\left[f(X) - f(\mathbb{E}X)\right] = \inf_{a\in I} \mathbb{E}\left[f(X) - f(a) - f'(a)(X-a)\right]$$

P.2 Sea Y una variable no negativa tal que $\mathbb{E}\left[Y \log Y\right] < \infty$. Mostrar que

$$\mathrm{Ent}(Y) = \inf_{u>0} \mathbb{E}\left[Y(\log Y - \log u) - (Y-u)\right].$$

P.3 Sea Z_i una función de las variables en $X^{(i)}$. Mostrar que

$$\mathrm{Ent}^{(i)}\left(e^{\lambda Z}\right) \leq \mathbb{E}^{(i)}\left[e^{\lambda Z}\phi(-\lambda(Z-Z_i))\right].$$

P.4 Mostrar que

$$\mathrm{Ent}\left(e^{\lambda Z}\right) \leq \sum_{i=1}^n \mathbb{E}\left[e^{\lambda Z}\phi(-\lambda(Z-Z_i))\right].$$

P.5 Justificar que $\forall\lambda \in \mathbb{R}$ y $\forall u \in [0,1]$, $\phi(-\lambda u) \leq u\phi(-\lambda)$. Sea $Z = f(X_1,\dots,X_n)$ donde f es acotada por si misma. Deducir la desigualdad diferencial

$$\left(\frac{\psi_{Z-\mathbb{E}Z}(\lambda)}{e^\lambda - 1}\right)' \leq \mathbb{E}Z \cdot \left(\frac{-\lambda}{e^\lambda - 1}\right)'$$

P.6 Mostrar que $\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \phi(\lambda)\mathbb{E}Z$ y que $\mathbb{P}\left(Z \geq \mathbb{E}Z + t\right) \leq \exp\left(-\frac{t^2}{2\mathbb{E}Z+2t/3}\right)$

Una variable definida por una función acotada por si misma se **concentra**.

Preguntas opcionales :

Una propiedad Π *definida sobre una union finita de productos* de un conjunto $\mathcal{X}$ es una secuencia $\Pi_1, \dots, \Pi_n$ tal que $\Pi_1 \subset \mathcal{X}, \dots, \Pi_n \subset \mathcal{X}^n$. Digamos que la $(x_1, \dots, x_m) \in \mathcal{X}^m$ satisface la propiedad Π si $(x_1, \dots, x_m) \in \Pi_m$. Una propiedad es *hereditaria* si por cada secuencia $(x_1, \dots, x_m)$ que satisfaga la propiedad Π cada sub-secuencia $(x_{i_1}, \dots, x_{i_k})$ satisface Π.

P'.1 (Ejercicios Hora 3) Sea f acotada por si misma y $Z = f(X_1, \dots, X_n)$ donde los X_i son variables aleatorias independientes. Mostrar que $\mathrm{Var}(Z) \leq \mathbb{E}Z$.

P'.2 Sea Π una propiedad hereditaria. Para cada $(x_1, \dots, x_m)$, se asocia el tamaño máximo de una sub-secuencia de $(x_1, \dots, x_m)$ que satisfaga Π. Denotamos $f_\Pi(x_1, \dots, x_m)$ este valor. Mostrar que f_Π es acotada por si misma.

Una función f tal que existe una propiedad Π tal que $f = f_\Pi$ se llama *función de configuración.*

P'.3 Sean $X_1, \dots, X_n$ i.i.d. discretas. Sea Z el numero de valores distintos que tomen las variables $X_1, \dots, X_n$. Mostrar que Z es una función de configuración de las $X_1, \dots, X_n$.

P'.4 (Utilizamos las notaciones de la tarea 2) Para una clase $\mathcal{A}$ de subconjuntos de $\mathbb{R}^d$ y elementos $x_1, \dots, x_n \in \mathbb{R}^d$, digamos que $\mathcal{A}$ rompe $(x_1, \dots, x_n)$ si $|\mathcal{A}(x_1^n)| = 2^n$. Denotamos $VC(\mathcal{A}, x_1^n)$ el tamaño máximo de una sub-secuencia de $(x_1, \dots, x_n)$ que está rota por $\mathcal{A}$. Mostrar que este noción de dimensión VC es una función de configuración.

19.4 Tarea 4

El objectivo de esta tarea es investigar condiciones sobre distribuciones para que éstas sean definidas en forma única por sus momentos. Dada una medida μ, definimos para cada $p \in \mathbb{N}$, el momento de orden p como

$$\mu_p = \int x^p d\mu.$$

Denotamos como $\mathcal{M}(\mathbb{R})$ al conjunto de medidas de probabilidad sobre $\mathbb{R}$ que tienen momentos finitos de cualquier orden y $\mathcal{M}(\mathbb{R}_+)$ las medidas de probabilidad sobre $\mathbb{R}_+$ que tienen momentos finitos de cualquier orden. Dada una sucesión real $\mathbf{m} = (m_p)_{p\in\mathbb{N}}$, sean

$$K(\mathbf{m}) = \{\mu \in \mathcal{M}(\mathbb{R}) : \forall p \in \mathbb{N}, \mu_p = m_p\} \qquad \text{(Problema de Hamburger)}$$

y

$$K^+(\mathbf{m}) = \{\mu \in \mathcal{M}(\mathbb{R}_+) : \forall p \in \mathbb{N}, \mu_p = m_p\} \qquad \text{(Problema de Stieljes)}$$

las soluciones de los problemas de momentos de Hamburger y de Stieljes. Decimos que una variable aleatoria X con sucesión de momentos $\mathbf{m}$ es *únicamente definida por sus momentos (UDM)* si $K(\mathbf{m})$ tiene un único elemento.

P.1 Mostrar que una variable aleatoria X tal que $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$ y $\mathbb{E}[X^4] = 1$ es igual (en distribución) a una variable de Rademacher (que vale 1 con probabilidad $\frac{1}{2}$ y -1 con probabilidad $\frac{1}{2}$). Dar un ejemplo de secuencia $\mathbf{m}$ tal que $K(\mathbf{m}) = \emptyset$.

En el siguiente, suponemos que $\mathbf{m}$ es tal que $K(\mathbf{m}) \neq \emptyset$.

P.2 Sea X una variable aleatoria de medida $\mu \in \mathcal{M}(\mathbb{R})$ de soporte finito y sea $\mathbf{m}$ la sucesión de los momentos μ_p de μ. Mostrar que X es UDM. *Pista : Para mostrar que* $\sum z_i x_i^p = 0$ *implica* $\forall i, z_i = 0$ *considere una formulación matricial y use el determinante de Vandermonde.*
Mostrar que, efectivamente, X es UDM $\mu_1, \mu_2, \ldots, \mu_{2n}$ donde n es el numero de átomos de μ.

P.3 Suponemos que μ es de soporte compacto. Usar el teorema de Portmanteau para probar que X es UDM.

P.4 Una variable beta (α, β) es una variable aleatoria de densidad sobre $[0, 1]$ igual a

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Calcular los momentos de una variable beta (α, β). ¿Qué podemos decir de una sucesión $(X_n)_n$ de variables aleatorias tal que para cada $p \in \mathbb{N}$ tenemos

$$\mathbb{E}[X_n^p] \underset{n\to+\infty}{\longrightarrow} \prod_{i=0}^{p-1} \frac{\alpha+i}{\alpha+\beta+i}?$$

Recordamos el siguiente teorema :

Teorema Una función holomorfa en un abierto U que vale 0 sobre un conjunto que tiene un punto de acumulación de U es nula sobre todo U. *Los interesados pueden encontrar el teorema y su prueba en el libro: Real and Complex analysis, 1987, W. Rudin*

P.5.a Mostrar que para cada $w > 0$ y cada $z \in \mathbb{C}$ tal que $\Re(z) > 0$ (parte real es positiva),

$$\int_0^{+\infty} t^{w-1} e^{-zt} dt = \frac{1}{z^w}\Gamma(w).$$

P.5.b Sea $g : \mathbb{R} \mapsto \mathbb{R}$,

$$g(x) = \frac{2}{3\sqrt{\pi}} |x|^{-2/3} \exp(-|x|^{2/3}) \cos\left(\frac{\pi}{3} + \sqrt{3}|x|^{2/3}\right).$$

Mostrar que para cada $p \in \mathbb{N}$,

$$\int_{-\infty}^{+\infty} x^p g(x) dx = 0.$$

P.5.c Sea $f : \mathbb{R} \mapsto \mathbb{R}$ la función de densidad

$$f(x) = \frac{1}{3\sqrt{\pi}}|x|^{-2/3}\exp(-|x|^{2/3}).$$

Mostrar que para $\rho \in [0, 1/2]$, $f + \rho g$ es una función de densidad y que para cada $p \in \mathbb{N}$,

$$\int_{-\infty}^{+\infty} x^p f(x)dx = \int_{-\infty}^{+\infty} x^p (f + \rho g)(x)dx$$

Deducir que X, cuya densidad es f, no es UDM.

P.5.d Sea $\mathbf{m}$ tal que si p es par $m_p = 0$ y si p es impar $m_p = (3p-1)(3p-3)\dots 1$. ¿Qué podemos decir de $K(\mathbf{m})$?

P.6.a Mostrar que $K(\mathbf{m})$ es un conjunto convexo.

P.6.b Mostrar que $K(\mathbf{m})$ es un conjunto compacto.

P.7 Mostrar lo siguiente:

Proposición Sea X una variable aleatoria sobre $\mathbb{R}$ de medida de probabilidad $\mu \in \mathcal{M}(\mathbb{R})$. Suponemos que la serie de Laplace

$$\sum_{p \geq 1} \frac{\mu_p}{p!} z^p$$

es de radio de convergencia no nulo, entonces X es UDM.

P.8.a Calcular el radio de convergencia de la serie de Laplace para X una variable $\mathcal{N}(0, 1)$. Deducir que X es UDM.

P.8.b Calcular los momentos de $Y = \exp(N)$, donde $N \sim \mathcal{N}(0, 1)$. ¿Qué el radio de convergencia de la serie de Laplace?

Admitimos el teorema siguiente:

Teorema Sea X una variable de densidad f positiva sobre $\mathbb{R}_+$. Si

$$\int_0^{+\infty} \frac{-\log f(t)}{1+t^2} dt < +\infty,$$

entonces X no es UDM.

P.9.a Mostrar que la variable aleatoria Y definida en **P.8.b** no es UDM.

P.9.b Sea $Z = W^3$ donde W es una variable aleatoria exponencial de parametro 1. Mostrar que Z no es UDM.

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] Luc Devroye and Gabor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media, 2001.

[6] 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.

[7] 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.

[8] 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.

[9] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

[10] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[11] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 2006.

[12] 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.

[13] Endre Szemerédi. Regular partitions of graphs. Technical report, Stanford University, 1975.

[14] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[15] Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[16] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[17] V.N. Vapnik and A. Ya. Chervonenrkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

[18] David Williams. *Probability with martingales*. Cambridge university press, 1991.