Chapter 1

Probability review

1.1 Measure theory for probabilists

In this first section, we recall the main notions of measure theory needed in probability theory. Standard (advanced) textbooks on the subject are [5, 8]. A first introduction is given in [43].

1.1.1 Probability space

Definition 1.1.1. *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a given set (sometimes called the "fundamental set"), $\mathcal{F}$ is a σ-field on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. We recall here that*

- A σ-field (or σ-algebra) on Ω is a collection $\mathcal{F}$ of subsets of Ω such that
(i) $\emptyset \in \mathcal{F}$,
(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,
(iii) if $(A_n,\, n \geq 1) \subset \mathcal{F}$, then $\cup_{n\geq 1} A_n \in \mathcal{F}$.

- A probability measure (or probability distribution, or distribution) on $(\Omega, \mathcal{F})$ is an application $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that
(i) $\mathbb{P}(\emptyset) = 0$,
(ii) if $(A_n,\, n \geq 1) \subset \mathcal{F}$ is such that $A_n \cap A_m = \emptyset$ for all $n \neq m$, then

$$\mathbb{P}\left(\cup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n),$$

(iii) $\mathbb{P}(\Omega) = 1$.

Definition 1.1.2. *A sub-σ-field of $\mathcal{F}$ is a collection $\mathcal{G}$ of subsets of Ω such that*
(i) $\mathcal{G} \subset \mathcal{F}$ (that is, if $A \in \mathcal{G}$, then $A \in \mathcal{F}$),
(ii) $\mathcal{G}$ is itself a σ-field.

Definition 1.1.3. *Given a collection $\mathcal{A}$ of subsets of Ω, the σ-field generated by $\mathcal{A}$ and denoted by $\sigma(\mathcal{A})$ is the smallest σ-field on Ω that contains $\mathcal{A}$.*

Example 1.1.4. The σ-field on $\mathbb{R}$ generated by the collection of open intervals $]a, b[$ with $a < b$, is called the Borel σ-field on $\mathbb{R}$ and is denoted by $\mathcal{B}(\mathbb{R})$. Elements of $\mathcal{B}(\mathbb{R})$ are called Borel sets. Note that $\mathcal{B}(\mathbb{R})$ is also generated by the collection of semi-infinite intervals $]-\infty, x]$, $x \in \mathbb{R}$.

Example 1.1.5. The measure μ that assigns to each interval $]a, b[\subset \mathbb{R}$ its length $b - a$ can be extended uniquely to all subsets of $\mathcal{B}(\mathbb{R})$. It is called the Lebesgue measure on $\mathbb{R}$ and and

is denoted by $\mu(B) = |B|$. Note that it is not a probability measure, since $\mu(\mathbb{R}) = \infty$, but it verifies points (i) and (ii) of the definition. Moreover, it becomes a probability measure when restricted to the interval $[0, 1]$.

Remark 1.1.6. In this course, probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ will generally be denoted by the letter μ, in order to avoid confusion with probability measures $\mathbb{P}$ on general spaces $(\Omega, \mathcal{F})$.

1.1.2 Random variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 1.1.7. *A random variable on $(\Omega, \mathcal{F})$ is an application $X : \Omega \to \mathbb{R}$ such that*

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Note that in order to check that a given application X is a random variable, it is sufficient to verify that

$$\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}, \quad \forall t \in \mathbb{R}.$$

Remark 1.1.8. The set $\{\omega \in \Omega : X(\omega) \in B\}$ is often simply denoted by $\{X \in B\}$.

Definition 1.1.9. *- Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A random variable X is said to be $\mathcal{G}$-measurable if*

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{G}, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

- $\sigma(X)$ denotes the σ-field generated by the collection of subsets $\{\{X \in B\}, B \in \mathcal{B}(\mathbb{R})\}$ (note that this collection of subsets is itself a σ-field; it is therefore equal to the σ-field $\sigma(X)$).

Remark 1.1.10. - X is a $\mathcal{G}$-measurable random variable if and only if $\sigma(X) \subset \mathcal{G}$.
- A $\mathcal{B}(\mathbb{R})$-measurable random variable $f : \mathbb{R} \to \mathbb{R}$ is also called a Borel-measurable function.
- Any continuous function $f : \mathbb{R} \to \mathbb{R}$ is Borel-measurable.
- If X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function, then $f(X)$ is a random variable.

Definition 1.1.11. *The law (or distribution) of a random random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the probability measure μ_X defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as*

$$\mu_X(B) = \mathbb{P}(\{X \in B\}), \quad B \in \mathcal{B}(\mathbb{R}).$$

A random variable (considered separately) is essentially characterized by its law. Two random variables X, Y with the same law are said to be identically distributed (i.d.) and this is denoted by $X \sim Y$. We distinguish two particular types of random variables.

A) *Discrete* random variables. If X takes its values in a countable set D, then its law μ_X is entirely characterized by the sequence of non-negative numbers $(\mu_X(\{x\}), x \in D)$ that sums up to 1.

B) *Continuous* random variables. If the law of X is absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}$ (this is to say that $\mu_X(B)=0$ for all Borel sets B such that $|B| = 0$), then the Radon-Nikodym theorem implies that there exists a Borel-measurable function $p_X : \mathbb{R} \to \mathbb{R}_+$, called the density function of X, such that

$$\mu_X(B) = \int_B p_X(x)\, dx, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

This implies in particular that $\int_{\mathbb{R}} p_X(x)\, dx = 1$.

Remark 1.1.12. In the expression $p_X(x)$, the subscript X recalls that the application p_X is the density function of the variable X (as it is the case for the law μ_X), whereas x is a real number, namely the point in which the application p_X is evaluated.

Definition 1.1.13. *For $A \in \mathcal{F}$, the application $1_A : \Omega \to \mathbb{R}$, called the indicator function of A, is defined as*

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

It is a random variable and is moreover discrete (taking only two values 0 and 1).

1.1.3 Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that an event $A \in \mathcal{F}$ is said to be negligible if $\mathbb{P}(A) = 0$. On the contrary, an event $B \in \mathcal{F}$ is said to be almost sure (a.s.) if $\mathbb{P}(B) = 1$.

The definition of the expectation (or average, or mean, or even Lebesgue's integral, depending on the context) of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is made in three steps.

Step 1. For any random variable X of the form

$$X(\omega) = \sum_{i=0}^{\infty} x_i \, 1_{A_i},$$

where x_i are non-negative numbers and $A_i \in \mathcal{F}$, we define

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} x_i \, \mathbb{P}(A_i) \in [0, \infty].$$

In particular, note that $\mathbb{E}(1_A) = \mathbb{P}(A)$.

Step 2. For any non-negative random variable X, let us define the following sequence of random variables

$$X_n(\omega) = \sum_{i=0}^{\infty} \frac{i}{2^n} \, 1_{\{\frac{i}{2^n} \le X < \frac{i+1}{2^n}\}}(\omega).$$

Note that X_n is of the type defined above: $\frac{i}{2^n} \ge 0$ and $A_i = \{\frac{i}{2^n} \le X < \frac{i+1}{2^n}\} \in \mathcal{F}$ because X is $\mathcal{F}$-measurable. Moreover, $X_n(\omega) \le X_{n+1}(\omega)$ for all n and ω and $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ (X_n is a sequence of staircases with more and more refined steps below the function X). It is therefore natural to define

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n) = \lim_{n\to\infty} \sum_{i=0}^{\infty} \frac{i}{2^n} \, \mathbb{P}(\{\frac{i}{2^n} \le X < \frac{i+1}{2^n}\}) \in [0, \infty].$$

Note that the sequence $(\mathbb{E}(X_n))$ is also non-decreasing and therefore converging in $[0, \infty]$.

Step 3. For any random variable X, let us define $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$, its positive and negative parts. Note that $|X| = X^+ + X^- \ge 0$. We say that X is integrable if $\mathbb{E}(|X|) < \infty$ and we define

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Another notation for $\mathbb{E}(X)$ is the following:

$$\int_\Omega X(\omega) \, d\mathbb{P}(\omega) \quad \left(\text{or } \int_\Omega X(\omega) \, \mathbb{P}(d\omega) \right).$$

When $\mathbb{P} = \mu$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $X = f$ is a Borel-measurable function, this notation becomes

$$\int_{\mathbb{R}} f(x)\, d\mu(x).$$

We have the following useful formulas for computing expectations.

Proposition 1.1.14. *Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}(|g(X)|) < \infty$. Then*

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\, d\mu_X(x),$$

where μ_X is the law of X. We moreover have the following two particular cases.

A) If X is a discrete random variable with values in D, then

$$\mathbb{E}(g(X)) = \sum_{x \in D} g(x)\, \mathbb{P}(\{X = x\}).$$

B) If X is a continuous random variable with density function p_X, then

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\, p_X(x)\, dx.$$

We list hereafter some well known properties of expectation.

(i) Linearity: if $c \in \mathbb{R}$ and X, Y are integrable, then $\mathbb{E}(cX + Y) = c\,\mathbb{E}(X) + \mathbb{E}(Y)$.

(ii) Monotony: if X, Y are integrable and $X \geq Y$ a.s. (that is, $\mathbb{P}(\{X \geq Y\}) = 1$), then $\mathbb{E}(X) \geq \mathbb{E}(Y)$. This implies in particular positivity: if $X \geq 0$ a.s. then $\mathbb{E}(X) \geq 0$.

(iii) Strict positivity: if $X \geq 0$ a.s. and $\mathbb{E}(X) = 0$, then $X = 0$ a.s.

Furthermore, we say that a random variable X is square-integrable if $\mathbb{E}(X^2) < \infty$ and that it is bounded if there exists $K > 0$ such that $|X| \leq K$ a.s. We have the following series of implications:

$$X \text{ is bounded} \quad \Rightarrow \quad X \text{ is square-integrable} \quad \Rightarrow \quad X \text{ is integrable,}$$

$$X \text{ is integrable and } Y \text{ is bounded} \quad \Rightarrow \quad XY \text{ is integrable}$$

and

$$X, Y \text{ are both square-integrable} \quad \Rightarrow \quad XY \text{ is integrable.}$$

We define the variance of a square-integrable random variable X as

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0$$

and the covariance of two square-integrable random variables X, Y as

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

1.1.4 Inequalities

We list here some of the well known inequalities concerning expectation.

Proposition 1.1.15. *(Cauchy-Schwarz' inequality)*
If X, Y are square-integrable random variables, then

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)}\,\sqrt{\mathbb{E}(Y^2)}.$$

Proposition 1.1.16. *(Chebychev's inequality)*
If X is a random variable and $\varphi : \mathbb{R} \to \mathbb{R}_+$ is increasing on $\mathbb{R}_+$ and such that $\mathbb{E}(\varphi(X)) < \infty$, then for any $a > 0$, we have

$$\mathbb{P}(\{X > a\}) \leq \frac{\mathbb{E}(\varphi(X))}{\varphi(a)}.$$

Note that the above inequality has actually different names (Chebychev, Markov, Bernstein, Chernoff, ...), depending on the community of researchers. We shall use it often with $\varphi(x) = \exp(tx)$, $t > 0$.

Proposition 1.1.17. *(Jensen's inequality)*
If X is a random variable and $\psi : \mathbb{R} \to \mathbb{R}$ is convex and such that $\mathbb{E}(|\psi(X)|) < \infty$, then

$$\psi(\mathbb{E}(X)) \leq \mathbb{E}(\psi(X)).$$

1.1.5 Convergence theorems

Let us cite here the three famous convergence theorems of measure theory.

Lemma 1.1.18. *(Fatou's lemma)*
If (X_n) is a sequence of non-negative random variables, then

$$\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

Theorem 1.1.19. *(Beppo-Levi's monotone convergence theorem)*
If (X_n) is a sequence of non-negative random variables such that $X_n \leq X_{n+1}$ a.s. for all n, then

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}\left(\lim_{n\to\infty} X_n\right).$$

Theorem 1.1.20. *(Lebesgue's dominated convergence theorem)*
If (X_n) is a sequence of random variables such that $X = \lim_{n\to\infty} X_n$ exists a.s. and there exists an integrable random variable Y with $|X_n| \leq Y$ a.s. for all n, then

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X).$$

1.1.6 Independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We define the notion of independence for three different objects.

1) Independence of *events*. A collection $(A_1, ..., A_n)$ of events in $\mathcal{F}$ is said to be independent if

$$\mathbb{P}(A_1^* \cap \ldots \cap A_n^*) = \mathbb{P}(A_1^*) \cdots \mathbb{P}(A_n^*)$$

for any $A_i^* = A_i$ or A_i^c.

2) Independence of σ-*fields*. A collection $(\mathcal{G}_1, ..., \mathcal{G}_n)$ of sub-σ-fields of $\mathcal{F}$ is said to be independent if

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$$

for all $A_i \in \mathcal{G}_i$.

Remark 1.1.21. The collection of events $(A_1, ..., A_n)$ is independent if and only if the collection of σ-fields $(\sigma(A_1), ..., \sigma(A_n))$ is independent.

3) Independence of *random variables*. A collection $(X_1, ..., X_n)$ of random variables is said to be independent of the corresponding collection of σ-fields $(\sigma(X_1), ..., \sigma(X_n))$ is independent.

Remark 1.1.22. In order to check that the collection $(X_1, ..., X_n)$ is independent, it is actually sufficient to verify that

$$\mathbb{P}(X_1 \leq t_1, ..., X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \cdots \mathbb{P}(X_n \leq t_n),$$

for all $t_1, ..., t_n \in \mathbb{R}$.

Remark 1.1.23. - The independence of two objects is often denoted by the sign $\perp$; we may therefore write $\mathcal{G} \perp \mathcal{H}$, $X \perp Y$ or $X \perp \mathcal{G}$. Note that if $X \perp Y$ and $f, g : \mathbb{R} \to \mathbb{R}$ are Borel-measurable functions, then $f(X) \perp g(Y)$.

1.2 Conditional expectation

The conditional probability of an event A given an event B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{given that } \mathbb{P}(B) > 0.$$

In a similar way, we may define

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X\, 1_B)}{\mathbb{P}(B)}, \quad \text{given that } \mathbb{P}(B) > 0.$$

This generalizes easily to conditioning with respect to a discrete random variable Y with values in a countable set D:

$$\begin{aligned} \mathbb{P}(A|Y) &= \varphi(Y), \qquad \text{where } \varphi(y) = \mathbb{P}(A|Y = y), \quad y \in D. \\ \mathbb{E}(X|Y) &= \psi(Y), \qquad \text{where } \psi(y) = \mathbb{E}(X|Y = y), \quad y \in D. \end{aligned}$$

However, how can we generalize such a formula for a continuous random variable Y, since $\mathbb{P}(Y = y) = 0$ for all $y \in \mathbb{R}$? The answer is obtained by conditioning with respect to a σ-field.

Definition 1.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The conditional expectation of X given $\mathcal{G}$ is the random variable Z such that*

(i) Z is $\mathcal{G}$-measurable,

(ii) $\mathbb{E}(ZY) = \mathbb{E}(XY)$ for any random variable Y $\mathcal{G}$-measurable and bounded.

Z is denoted by $\mathbb{E}(X|\mathcal{G})$.

Remark 1.2.2. The existence of Z is guaranteed by Radon-Nikodym's theorem and it follows from the definition that Z is integrable. Note moreover that if both Z_1, Z_2 satisfy (i) and (ii), then $Z_1 = Z_2$ a.s., so the conditional expectation is well defined up to a negligible set.

We further define
- $\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(1_A|\mathcal{G})$ for $A \in \mathcal{F}$.
- $\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y))$ for a random variable Y.

Remark 1.2.3. Note that $\mathbb{E}(X|\mathcal{G})$ or $\mathbb{E}(X|Y)$ is a *random variable*, whereas in many textbooks on information theory, conditioning often denotes the expectation of some random variable; think for example at the conditional entropy of two random variables with joint density function $p_{X,Y}$:

$$h(X|Y) = -\mathbb{E}(\log(p_{X|Y})) = -\mathbb{E}(\log(p_{X,Y})) + \mathbb{E}(\log(p_Y)).$$

(recall that $p_{X|Y} = \frac{p_{X,Y}}{p_Y}$ by definition).

Example 1.2.4. When do we have an explicit expression for conditional expectation? This is the case at least in two particular situations, for which we recover classical formulas.

A) If X, Y are two discrete random variables with values in a countable set D, then

$$E(X|Y) = \psi(Y), \quad \text{where } \psi(y) = \sum_{x \in D} x\, \mathbb{P}(X = x|Y = y), \quad y \in D.$$

B) If X, Y are two continuous random variables with joint density function $p_{X,Y}$, then

$$E(X|Y) = \psi(Y), \quad \text{where } \psi(y) = \int_{\mathbb{R}} x\, \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dy, \quad y \in \mathbb{R},$$

and p_Y is the marginal density function of Y given by $p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x,y)dy$, assumed here to be positive everywhere for simplicity.

In many other situations however, the conditional expectation of a random variable with respect to some general σ-field is not directly computable. We therefore need some rules in order to proceed; these are listed below.

1) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
2) If X is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s.
3) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.
4) If Y is $\mathcal{G}$-measurable and bounded, then $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})\, Y$ a.s.
5) If $\mathcal{H}$ is a sub-σ-field of $\mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ a.s.

Moreover, Jensen's inequality is also valid for conditional expectation.

Proposition 1.2.5. *Let X be a random variable, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be convex and such that $\mathbb{E}(|\psi(X)|) < \infty$. Then*

$$\psi(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(\psi(X)|\mathcal{G}) \quad a.s.$$

A further property is given in the following proposition.

Proposition 1.2.6. *Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$, X, Y be two random variables such that X is independent of $\mathcal{G}$ and Y is $\mathcal{G}$-measurable, an let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}(|\varphi(X,Y)|) < \infty$. Then*

$$\mathbb{E}(\varphi(X,Y)|\mathcal{G}) = \psi(Y), \quad \textit{where } \psi(y) = \mathbb{E}(\varphi(X,y)).$$

This property has the following consequence: when computing the expectation of a function φ of two independent random variables X and Y, one can always divide the computation in two steps by writing

$$\mathbb{E}(\varphi(X,Y)) = \mathbb{E}(\mathbb{E}(\varphi(X,Y)|Y)) = \mathbb{E}(\psi(Y))$$

where $\psi(y) = \mathbb{E}(\varphi(X,y))$ (this is actually nothing but Fubini's theorem).

Homework 1.2.7. Using only definition 1.2.1, prove
- formulas A) and B) in example 1.2.4.
- properties 1) to 5).
- proposition 1.2.6 in the case where X, Y are discrete and $\mathcal{G} = \sigma(Y)$.

1.3 Convergences of sequences of random variables

For a given sequence of random variables $(X_n,\, n \geq 1)$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there are several notions of convergence to a limiting random variable X. Let us review the most important ones.

The first one is convergence in probability.

Definition 1.3.1. *The sequence (X_n) is said to converge in probability to X (and this is denoted by $X_n \xrightarrow{\mathbb{P}} X$) if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

A stronger notion of convergence is that of almost sure convergence.

Definition 1.3.2. *The sequence (X_n) is said to converge almost surely to X (and this is denoted by $X_n \to X$ a.s.) if*

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1.$$

Almost sure convergence implies convergence in probability, whereas convergence in probability only implies that there exists a subsequence converging almost surely. Moreover, we have the following equivalent criterion for almost sure convergence:

$$X_n \to X \text{ a.s.} \quad \text{iff} \quad \forall \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0, \tag{1.3.1}$$

where "i.o." stands for "infinitely often", that is, "for an infinite number of n".

Example 1.3.3. (Convergence in probability does not imply almost sure convergence)
Let (ξ_n) be a sequence of i.i.d. (that is, independent and identically distributed) random variables such that $\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = 0) = 1/2$. Let us define $X_1 = 1$, $X_2 = \xi_1$, $X_3 = 1 - \xi_1$, $X_4 = \xi_1\xi_2$, $X_5 = \xi_1(1-\xi_2)$, $X_6 = (1-\xi_1)\xi_2$, $X_7 = (1-\xi_1)(1-\xi_2)$, $X_8 = \xi_1\xi_2\xi_3$ and so on. It is easy to see that (X_n) is a sequence of 0's and 1's such that

$$\mathbb{P}(X_n = 1) = \frac{1}{2^j}, \quad \forall n \in \{2^j, ..., 2^{j+1} - 1\}.$$

It therefore converges to 0 in probability as $n \to \infty$, but not almost surely, since each realization of the sequence contains an infinite number of 1's.

A sufficient condition guaranteeing almost sure convergence is given in the following lemma.

Lemma 1.3.4. *(Borel-Cantelli)*
Let (A_n) be a sequence of events in $\mathcal{F}$.
a) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$.
b) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and the events (A_n) are independent, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$.

Let us recall here that $\limsup_{n\to\infty} A_n = \bigcap_{m\geq 1} \bigcup_{n\geq m} A_n$ and that

$$\omega \in \limsup_{n\to\infty} A_n \quad \text{iff} \quad \omega \in A_n \text{ } i.o. \tag{1.3.2}$$

Therefore, using (1.3.1) and part a) of the Borel-Cantelli lemma, we see that $X_n \to X$ a.s. if for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty.$$

Let us finally give two more notions of convergence for sequences of random variables.

Definition 1.3.5. *Let (X_n) be a sequence of integrable (resp. square-integrable) random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (X_n) is said to converge in mean (resp. quadratically) to X if*

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0 \quad (\textit{resp.} \lim_{n\to\infty} \mathbb{E}(|X_n - X|^2) = 0).$$

Note that by Chebychev's inequality, either convergence in mean or quadratic convergence implies convergence in probability.

Homework 1.3.6. - Prove (1.3.2).
- For a sequence (A_n) of events in $\mathcal{F}$, one also defines $\liminf_{n\to\infty} A_n = \bigcup_{m\geq 1} \bigcap_{n\geq m} A_n$. Note that

$$\omega \in \liminf_{n\to\infty} A_n \quad \text{iff there exists } m \geq 1 \text{ such that } \omega \in A_n \text{ for all } n \geq m, \tag{1.3.3}$$

and that

$$\left(\limsup_{n\to\infty} A_n\right)^c = \liminf_{n\to\infty} A_n^c. \tag{1.3.4}$$

Deduce (1.3.1) from (1.3.2), (1.3.3) and/or (1.3.4).

1.3.1 Laws of large numbers

Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let us define the sequence of empirical means

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad n \geq 1.$$

We recall below the two laws of large numbers.

Theorem 1.3.7. *(Weak law of large numbers)*
a) If $\mathbb{E}(|X_1|) < \infty$, then

$$S_n \xrightarrow{\mathbb{P}} \mathbb{E}(X_1).$$

b) If $\lim_{a\to\infty} a\, \mathbb{P}(|X_1| > a) = 0$, then

$$S_n - \mathbb{E}(X_1\, 1_{\{|X_1|\leq n\}}) \xrightarrow{\mathbb{P}} 0.$$

Remark 1.3.8. - Assumptions are made only on X_1, since the X_n are i.i.d.
- Assumption in part b) does not imply that $\mathbb{E}(|X_1|) < \infty$. It is actually the weakest condition under which convergence in probability takes place.

Theorem 1.3.9. *(Strong law of large numbers)*
a) If $\mathbb{E}(|X_1|) < \infty$, *then*

$$S_n \to \mathbb{E}(X_1) \quad a.s.$$

b) If $\mathbb{E}(|X_1|) = \infty$, *then*

$$\limsup_{n\to\infty} |S_n| = \infty \quad a.s., \quad \text{that is, } (S_n) \text{ diverges a.s.}$$

The conclusion of the strong law (part a) is definitely stronger than that of the weak law (part a). One can however appreciate the difference of the conclusions when the assumption $\mathbb{E}(|X_1|) < \infty$ is relaxed.

Homework 1.3.10. - Show that

$$\mathbb{E}(|X_1|) = \int_0^\infty \mathbb{P}(|X_1| > a)\, da$$

and use this to prove that in the weak law, assumption a) implies assumption b).
- Prove the weak law (part a) under the assumption that $\mathbb{E}(X_1^2) < \infty$.
- Prove the strong law (part a) under the assumption that $\mathbb{E}(X_1^4) < \infty$ and $\mathbb{E}(X_1) = 0$.

1.4 Distributions

Let μ be a distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In this section, we define a series of objects related to μ and discuss their properties.

1.4.1 Distribution function

Definition 1.4.1. *The distribution function of a distribution* μ *is the application* $F_\mu : \mathbb{R} \to [0,1]$ *defined as*

$$F_\mu(t) = \mu(\,]-\infty, t]), \quad t \in \mathbb{R}.$$

F_μ has the following properties:

(i) F_μ is non-decreasing on $\mathbb{R}$.
(ii) $\lim_{t\to-\infty} F_\mu(t) = 0$ and $\lim_{t\to+\infty} F_\mu(t) = 1$.
(iii) F_μ is right-continuous on $\mathbb{R}$, that is, $\lim_{\varepsilon\downarrow 0} F_\mu(t+\varepsilon) = F_\mu(t)$, for all $t \in \mathbb{R}$.

Reciprocally, we have the following proposition.

Proposition 1.4.2. *Any function* F *satisfying conditions (i) to (iii) is the distribution function of some distribution* μ, *and there is a one-to-one correspondence between the set of distributions and the set of distribution functions.*

Let us consider the following two particular cases.

A) If μ is discrete (that is, there exists a countable set D such that $\mu(D)$=1), then F_μ is the step function given by

$$F_\mu(t) = \sum_{x\in D,\, x\le t} \mu(\{x\}), \quad \forall t \in \mathbb{R}.$$

Remark 1.4.3. A frequent notation for discrete distributions is Dirac's notation:

$$\mu = \sum_{x \in D} \mu(\{x\})\, \delta_x,$$

where δ_x denotes the distribution with values in $\{0, 1\}$ defined as

$$\delta_x(B) = \begin{cases} 1, & \text{if } x \in B, \\ 0, & \text{if } x \notin B, \end{cases}$$

and whose distribution function is the step function $F_x(t) = 1_{[x,\infty[}(t)$, $t \in \mathbb{R}$ (also known as Heaviside's function).

B) If μ is absolutely continuous with respect to Lebesgue's measure (that is, $\mu(B)$=0 for all Borel sets B such that $|B| = 0$), then F_μ is continuous and differentiable, its derivative being the density function p_μ of μ and

$$F_\mu(t) = \int_{-\infty}^{t} p_\mu(x)\, dx, \quad \forall t \in \mathbb{R}.$$

Note that because p_μ is the derivative of F_μ, one sometimes finds the notation $p_\mu(x) = \frac{d\mu}{dx}(x)$.

In general, F_μ can be a combination of continuous and step functions, or even something more complicated such as the devil staircase illustrated below:

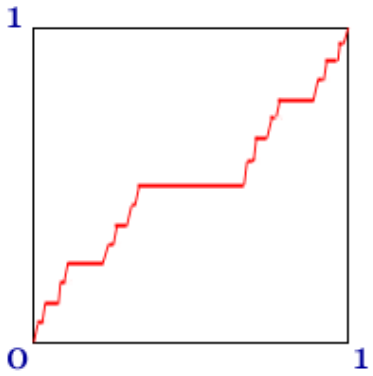

Figure 1.1: devil's staircase

This function has the following strange properties. First of all, the set where it is constant has Lebesgue's measure equal to 1 (so the function is constant almost everywhere), but on the other hand, the function is continuous and increasing from 0 to 1! It is neither a step function (being continuous), nor an absolutely continuous function (because its density function would be equal to zero almost everywhere, which is impossible).

Homework 1.4.4. - Prove properties (i) to (iii).
- Give an example of distribution μ such that F_μ is not left-continuous.
- From figure 1, deduce a mathematical definition of the devil staircase and prove the assertions made about it.

1.4.2 Characteristic function

The characteristic function is an object of central importance for proving limit theorems concerning sums of independent random variables, as shown by proposition 1.4.7 below.

Definition 1.4.5. *The characteristic function (or Fourier transform) of a distribution μ is the application $\phi_\mu : \mathbb{R} \to \mathbb{C}$ defined as*

$$\phi_\mu(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x), \quad t \in \mathbb{R}.$$

ϕ_μ has the following properties:

(i) $\phi_\mu(0) = 1$.
(ii) ϕ_μ is continuous on $\mathbb{R}$.
(iii) ϕ_μ is non-negative definite, that is,

$$\sum_{j,k=1}^{n} c_j\, \overline{c_k}\, \phi_\mu(t_j - t_k) \geq 0, \quad \forall n \geq 1,\ c_1, ..., c_n \in \mathbb{C},\ t_1, ..., t_n \in \mathbb{R}.$$

Proof. (i) is clear.
(ii) Using the dominated convergence theorem, we see that

$$|\phi_\mu(t) - \phi_\mu(s)| \leq \int_{\mathbb{R}} |e^{itx} - e^{isx}|\, d\mu(x) \to 0, \quad \text{as} \quad |t - s| \to 0.$$

(iii) Let us simply compute

$$\sum_{j,k=1}^{n} c_j\, \overline{c_k}\, \phi_\mu(t_j - t_k) = \int_{\mathbb{R}} \sum_{j,k=1}^{n} c_j\, \overline{c_k}\, e^{it_j x}\, e^{-it_k x}\, d\mu(x) = \int_{\mathbb{R}} \left| \sum_{j=1}^{n} c_j\, e^{it_j x} \right|^2 d\mu(x) \geq 0.$$

□

Reciprocally, we have the following Fourier's inversion theorem.

Theorem 1.4.6. *Any function ϕ satisfying conditions (i) to (iii) is the characteristic function of some distribution μ, and there is a one-to-one correspondence between the set of distributions and the set of characteristic functions. Moreover, we have the following Fourier's inversion formula: for all $a < b$ continuity points of the distribution function F_μ,*

$$\mu(\,]a, b[\,) = F_\mu(b) - F_\mu(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_\mu(t)\, dt. \tag{1.4.1}$$

Furthermore, if $\int_{\mathbb{R}} |\phi_\mu(t)|\, dt < \infty$, then μ admits a bounded continuous density function p_μ given by

$$p_\mu(x) = F'_\mu(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx}\, \phi_\mu(t)\, dt, \quad x \in \mathbb{R}.$$

Proof. We only prove here inversion's formula 1.4.1. By definition of ϕ_μ and Fubini's theorem, we have for all $T > 0$,

$$\begin{aligned}
\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_\mu(t)\, dt &= \int_{\mathbb{R}} \left(\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{2\pi i t}\, dt \right) d\mu(x) \\
&= \int_{\mathbb{R}} \left(\int_{-T}^{T} \frac{\sin(t(x-a)) - \sin(t(x-b))}{2\pi t}\, dt - i \int_{-T}^{T} \frac{\cos(t(x-a)) - \cos(t(x-b))}{2\pi t}\, dt \right) d\mu(x)
\end{aligned}$$

Since $t \mapsto \frac{\cos(t(x-a))-\cos(t(x-b))}{2\pi t}$ is an odd function and

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{\sin(ct)}{2\pi t}\, dt = \begin{cases} \frac{1}{2}, & \text{if } c > 0, \\ -\frac{1}{2}, & \text{if } c < 0, \end{cases}$$

we obtain that

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{2\pi i t}\, dt = 1_{]a,b[}(x), \quad \forall x \neq a, b.$$

Since a, b are assumed to be continuity points of F_μ, the dominated convergence theorem finally implies that

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_\mu(t)\, dt = \int_{\mathbb{R}} 1_{]a,b[}(x)\, d\mu(x) = F_\mu(b) - F_\mu(a).$$

□

We list here some more properties of characteristic functions:

(iv) ϕ_μ is uniformly continuous on $\mathbb{R}$, that is, $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $|\phi_\mu(t) - \phi_\mu(s)| < \varepsilon$ whenever $|t - s| < \delta$.

(v) If μ is symmetric on $\mathbb{R}$ (that is, $\mu(B) = \mu(-B)$ for all $B \in \mathcal{B}(\mathbb{R})$), then $\phi_\mu(t) \in \mathbb{R}$, for all $t \in \mathbb{R}$.

(vi) If μ is compactly supported (that is, there exists $M > 0$ such that $\mu([-M, M]^c) = 0$), then ϕ_μ is C^∞ on $\mathbb{R}$; it is moreover the restriction to $i\mathbb{R}$ of the analytic function $\psi_\mu : \mathbb{C} \to \mathbb{C}$ defined by

$$\psi_\mu(z) = \int_{\mathbb{R}} e^{zx}\, d\mu(x), \quad z \in \mathbb{C}.$$

Note that ψ_μ restricted to $\mathbb{R}$ is Laplace's transform of μ.

Finally, let us recall the following result.

Proposition 1.4.7. *If μ, ν are two distributions on $\mathbb{R}$, then the characteristic function of their convolution product $\mu * \nu$ is given by*

$$\phi_{\mu*\nu}(t) = \phi_\mu(t)\, \phi_\nu(t), \quad t \in \mathbb{R}.$$

Since the convolution product $\mu * \nu$ is the law of the sum of two independent random variables X and Y with laws μ and ν respectively, this explains the importance of characteristic functions for analyzing sums of independent random variables. We will see an application of this in section 1.5.3.

Homework 1.4.8. Prove properties (iv) to (vi) and show that property (iii) implies that $\phi_\mu(-t) = \overline{\phi_\mu(t)}$ and that $|\phi_\mu(t)| \leq \phi_\mu(0) = 1$.

1.4.3 Moments

Definition 1.4.9. *The moment of order $k \geq 0$ of a distribution μ is the real number m_k defined as*

$$m_k = \int_{\mathbb{R}} x^k\, d\mu(x).$$

Note that m_j exists and is finite for all $j \leq k$ if and only if

$$\int_{\mathbb{R}} |x|^k \, d\mu(x) < \infty. \tag{1.4.2}$$

Given that all the moments $(m_k,\, k \geq 0)$ of a given distribution μ exist, they have the following properties:

(i) $m_0 = 1$.

(ii) The infinite matrix M whose entries are given by $M_{jk} = m_{j+k}$, $j, k \geq 0$ is non-negative definite, that is,

$$\sum_{j,k=1}^{n} c_j \, \overline{c_k} \, M_{jk} \geq 0, \quad \forall n \geq 1,\ c_1, ..., c_n \in \mathbb{C}.$$

Proof. (i) is clear.
(ii) Let us compute

$$\sum_{j,k=1}^{n} c_j \, \overline{c_k} \, m_{j+k} = \int_{\mathbb{R}} \sum_{j,k=1}^{n} c_j \, \overline{c_k} \, x^{j+k} \, d\mu(x) = \int_{\mathbb{R}} \left| \sum_{j=1}^{n} c_j \, x^j \right|^2 d\mu(x) \geq 0.$$

□

For a given sequence of moments (m_k), it is however not clear whether the underlying distribution μ is unique. In order to get an answer to this question, we first need to relate moments to characteristic functions.

(iii) Let $k \geq 1$. If (1.4.2) is satisfied, then ϕ_μ is k times continuously differentiable on $\mathbb{R}$ and

$$\frac{d^k \phi_\mu}{dt^k}\Big|_{t=0} = i^k \, m_k, \quad \text{so} \quad \phi_\mu(t) = \sum_{j=0}^{k} \frac{i^j \, m_j}{j!} \, t^j + o(t^k), \tag{1.4.3}$$

where $g(t) = o(t^k)$ means that $\lim_{t \to 0} \frac{|g(t)|}{|t|^k} = 0$. The first relation shows that ϕ_μ is a moment generating function.

(iv) If ϕ_μ is k times differentiable at 0, then

$$\int_{\mathbb{R}} |x|^{2p} \, d\mu(x) < \infty, \quad \forall p \in \mathbb{N} \text{ such that } 2p \leq k.$$

From these two properties, we deduce in particular that ϕ_μ is C^∞ on $\mathbb{R}$ if and only if all moments m_k exist. Even in this case however, many characteristic functions (and therefore many distributions) may correspond to a given sequence (m_k). The one-to-one correspondence is ensured by the following theorem.

Theorem 1.4.10. *If μ is a distribution such that its moments (m_k) satisfy*

$$\limsup_{k \to \infty} \frac{1}{2k} \, (m_{2k})^{\frac{1}{2k}} < \infty, \tag{1.4.4}$$

then μ is the unique distribution with sequence of moments (m_k).

Remark 1.4.11. Carleman showed that condition (1.4.4) can be replaced by the slightly weaker condition

$$\sum_{k=1}^{\infty} m_{2k}^{-\frac{1}{2k}} = \infty. \tag{1.4.5}$$

Both (1.4.4) and (1.4.5) are conditions limiting the growth of the sequence (m_{2k}). They may in turn be reformulated into conditions limiting the weight of the distribution's tail. Note that they are both satisfied if $m_{2k} \leq (Ck)^{2k}$ for some $C > 0$ (in which case the above series is greater than or equal to the harmonic series). Note moreover that these conditions only involve even moments m_{2k} because Cauchy-Schwarz' inequality guarantees that

$$m_{2k+1}^2 = \left(\int_{\mathbb{R}} x^k \, x^{k+1} \, d\mu(x) \right)^2 \leq \int_{\mathbb{R}} x^{2k} \, d\mu(x) \int_{\mathbb{R}} x^{2k+2} \, d\mu(x) = m_{2k} \, m_{2k+2}, \quad \forall k \geq 0. \tag{1.4.6}$$

Proof of theorem 1.4.10. (sketch)
We show the result under the much stronger condition that there exists $C > 0$ such that

$$m_{2k} \leq C^{2k}, \quad \forall k \geq 0. \tag{1.4.7}$$

(This condition is satisfied for example if we know a priori that the measure μ is compactly supported). By Cauchy's criterion, ϕ_μ admits the following Taylor's expansion

$$\phi_\mu(t) = \sum_{k \geq 0} \frac{i^k \, m_k}{k!} \, t^k, \quad |t| < R, \tag{1.4.8}$$

with convergence radius $R = \frac{1}{L}$, where

$$L = \limsup_{k \to \infty} \left(\frac{|m_k|}{k!} \right)^{\frac{1}{k}} = \limsup_{k \to \infty} \left(\frac{m_{2k}}{(2k)!} \right)^{\frac{1}{2k}},$$

using (1.4.6). Stirling's formula ($\log(k!) \sim k \log k$) then implies that

$$L = \limsup_{k \to \infty} \frac{1}{2k} \, (m_{2k})^{\frac{1}{2k}}$$

so condition (1.4.7) guarantees that $L = 0$, i.e. $R = \infty$, i.e. the function ϕ_μ is analytic on $\mathbb{R}$ (or more precisely, is the restriction to $\mathbb{R}$ of an analytic function on $\mathbb{C}$). It is therefore entirely determined by the sequence of its derivatives at 0, i.e. the sequence $(i^k \, m_k)$. Fourier's inversion's theorem 1.4.6 allows us to conclude that (m_k) determines μ entirely. □

Remark 1.4.12. Note that condition (1.4.4) only guarantees that $L < \infty$, i.e. $R > 0$. This does not allow us to conclude directly that ϕ_μ is uniquely determined by the sequence (m_k) (since it is not analytic outside $[-R, R]$), so the proof of the theorem becomes more delicate.

Example 1.4.13. A famous example of distribution which is not characterized by its moments is the log-normal distribution μ_0 with density function

$$p_0(x) = \frac{1}{\sqrt{2\pi}} \, \frac{1}{x} \, \exp\left(-\frac{(\log x)^2}{2} \right), \quad x > 0.$$

This distribution has moments $m_k = \exp(k^2/2)$ that do not satisfy condition (1.4.5) because

$$\sum_{k=1}^{\infty} (\exp(2k^2))^{-\frac{1}{2k}} = \sum_{k=1}^{\infty} \exp(-k) < \infty.$$

It can be checked that the distributions μ_a, $-1 \leq a \leq 1$, with corresponding densities

$$p_a(x) = p_0(x)\,(1 + a\sin(2\pi\log(x))), \quad x > 0,$$

have the same sequence of moments for any $-1 \leq a \leq 1$.

Homework 1.4.14. - Prove properties (iii) and (iv).
- Prove that any sequence of real numbers (m_k) satisfying (i), (ii) and condition (1.4.7) is indeed the sequence of moments of some distribution μ (hint: use characteristic function).
- Consider the sequence $(m_k = \frac{1}{k+1},\, k \geq 0)$. Is the related matrix M non-negative definite? What is then the associated distribution μ? Is this distribution unique?
- Check the assertions made in example 1.4.13 and show furthermore that for any $a > 0$, the *discrete* distributions ν_a defined by

$$\nu_a(\{ae^j\}) = c_a\, a^j\, \exp(-j^2/2), \quad j \in \mathbb{Z},$$

have the same moments $m_k = \exp(k^2/2)$ (where c_a is an appropriate normalization constant).
- For $\lambda > 0$, let μ_λ be the distribution with density function

$$p_\lambda(x) = c_\lambda\, \exp(-x^\lambda), \quad x > 0,$$

where c_λ is an appropriate normalization constant. For which values of λ is the distribution μ_λ uniquely determined by its moments? From this example, deduce approximately the limiting weight of a distribution's tail, above which condition (1.4.4) is not satisfied.

1.4.4 Stieltjes' transform

In random matrix theory, Stieltjes' transform plays a role similar to the one of characteristic function for sums of independent random variables.

Definition 1.4.15. *The Stieltjes (or Cauchy) transform of a distribution μ is the application $g_\mu : \mathbb{C} \to \mathbb{C}$ defined as*

$$g_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z}\, d\mu(x), \quad z \in \mathbb{C}.$$

Remark 1.4.16. g_μ is a priori ill-defined on $\mathbb{R}$ (or more precisely on supp μ). We will see however that the distribution μ is entirely determined by the behavior of g_μ on the set $\mathbb{C}_+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. The situation is much more complicated for measures μ with support in the complex plane. These appear when studying eigenvalues of non-Hermitian random matrices.

We immediately see from the definition that

$$|g_\mu(z)| \leq \int_{\mathbb{R}} \frac{1}{|x - z|}\, d\mu(x) \leq \int_{\mathbb{R}} \frac{1}{|\mathrm{Im}(z)|}\, d\mu(x) = \frac{1}{|\mathrm{Im}(z)|}.$$

Moreover, g_μ has the following properties:

(i) g_μ is analytic on $\mathbb{C}\backslash\mathbb{R}$ (it is actually analytic outside supp μ).
(ii) $\mathrm{Im}\,(g_\mu(z))\,\mathrm{Im}(z) > 0$ for all $z \in \mathbb{C}\backslash\mathbb{R}$.
(iii) $\lim_{v\to\infty} v\,|g_\mu(iv)| = 1$.

Proof. (i) For all $\varepsilon > 0$, the application $z \mapsto \frac{1}{x-z}$ is analytic on $\{z \in \mathbb{C} : |\mathrm{Im}(z)| \geq \varepsilon\}$, and the dominated convergence theorem allows us to show that the same is true for $g_\mu(z)$.
(ii) Denoting $z = u + iv$ with $v \neq 0$, we can decompose $g_\mu(z)$ into its real and imaginary parts:

$$g_\mu(u + iv) = \int_{\mathbb{R}} \frac{x - u}{(x - u)^2 + v^2}\, d\mu(x) + i \int_{\mathbb{R}} \frac{v}{(x - u)^2 + v^2}\, d\mu(x). \tag{1.4.9}$$

This implies that

$$\mathrm{Im}\,(g_\mu(z))\,\mathrm{Im}(z) = \int_{\mathbb{R}} \frac{v^2}{(x-u)^2+v^2}\,d\mu(x) > 0.$$

(iii) The above formula (1.4.9) gives also

$$v\,|\mathrm{Im}(g_\mu(iv)| = \left(\left(\int_{\mathbb{R}} \frac{xv}{x^2+v^2}\,d\mu(x)\right)^2 + \left(\int_{\mathbb{R}} \frac{v^2}{x^2+v^2}\,d\mu(x)\right)^2\right)^{\frac{1}{2}} \underset{v\to\infty}{\to} 1.$$

□

Reciprocally, we have the following proposition.

Proposition 1.4.17. *Any function g satisfying properties (i) to (iii) is the Stieltjes transform of some distribution μ, and there is a one-to-one correspondence between the set of distributions and the set of Stieltjes' transforms. Moreover, we have the following inversion formula: for all $a < b$ continuity points of the distribution function F_μ,*

$$\mu(\,]a,b[\,) = F_\mu(b) - F_\mu(a) = \lim_{v\downarrow 0} \frac{1}{\pi}\int_a^b Im\,(g_\mu(u+iv))\,du. \tag{1.4.10}$$

Furthermore, if μ admits a density p_μ, then

$$p_\mu(u) = F'_\mu(u) = \frac{1}{\pi}\lim_{v\downarrow 0} Im\,(g_\mu(u+iv)), \quad u\in\mathbb{R}.$$

Proof. We only prove here formula (1.4.10). Using (1.4.9) and Fubini's theorem, we obtain that for all $v > 0$,

$$\begin{aligned}\frac{1}{\pi}\int_a^b \mathrm{Im}\,(g_\mu(u+iv))\,du &= \frac{1}{\pi}\int_{\mathbb{R}}\left(\int_a^b \frac{v}{(x-u)^2+v^2}\,du\right)\,d\mu(x)\\ &= \int_{\mathbb{R}}\left(\frac{1}{\pi}\arctan\left(\frac{a-x}{v}\right) - \frac{1}{\pi}\arctan\left(\frac{b-x}{v}\right)\right)\,d\mu(x).\end{aligned}$$

Since the integrand converges to $1_{]a,b[}(x)$ for all $x \neq a, b$ as $v \downarrow 0$, and since a, b are continuity points of F_μ, the conclusion follows by dominated convergence theorem. □

Finally, let us consider the relation between Stieltjes' transform and moments in the case where μ is compactly supported (i.e. there exists $M > 0$ such that $\mu([-M,M]^c) = 0$ and all moments m_k exist). Using the following Taylor's expansion:

$$\frac{1}{x-z} = -\frac{1}{z(1-\frac{x}{z})} = -\frac{1}{z}\sum_{k\geq 0}\left(\frac{x}{z}\right)^k,$$

valid for all x such that $|x| < |z|$, we obtain the following Laurent's expansion of g_μ:

$$g_\mu(z) = -\frac{1}{z}\sum_{k\geq 0}\frac{m_k}{z^k}, \tag{1.4.11}$$

valid for $|z| > M$. One can deduce from this formula that $g_\mu(z)$ behaves like $-\frac{1}{z}$ as $|z| \to \infty$. Using Cauchy's formula, we moreover obtain that

$$m_k = \frac{i}{2\pi}\int_{C_R} z^k\,g_\mu(z)\,dz, \quad k \geq 0,$$

where C_R denotes the circle of radius $R > M$ centered at zero. In this sense, the Stieltjes transform is also a moment generating function.

Homework 1.4.18. - Compute the Stieltjes transform of the following distributions: $\mu = \delta_{x_0}$, $\mu = \mathcal{U}([0,1])$.
- Compute the inverse Stietjes' transform of

$$g_\mu(z) = \frac{-z + \sqrt{z^2 - 4}}{2}, \quad z \in \mathbb{C}.$$

Does μ admit a density function? How does it look like?

1.5 Weak convergence of sequences of distributions

This is the central notion of convergence for sequences of distributions. It has several equivalent formulations, which we review here.

Definition 1.5.1. *A sequence $(\mu_n, n \geq 1)$ of distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to converge weakly to a limiting distribution μ (and this is denoted by $\mu_n \Rightarrow \mu$) if for all $f \in C_b(\mathbb{R})$,*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\, d\mu_n(x) = \int_{\mathbb{R}} f(x)\, d\mu(x).$$

Remark 1.5.2. A distribution may be viewed as an element of the dual space of $C_b(\mathbb{R})$, the space of continuous bounded functions on $\mathbb{R}$. The term "weak convergence" comes from the fact that convergence of (μ_n) only takes place against test functions f, and not in some norm defined on the space of distributions. Note that from the strict point of view of functional analysis, this type of convergence should actually be called "weak-* convergence" and not "weak convergence".

We have the following characterization of weak convergence, known as portmanteau's theorem.

Theorem 1.5.3. *The following are equivalent:*
a) The sequence (μ_n) converges weakly to μ, according to definition 1.5.1.
b) For any closed set $G \subset \mathbb{R}$, $\limsup_{n\to\infty} \mu_n(G) = \mu(G)$.
c) For any open set $U \subset \mathbb{R}$, $\liminf_{n\to\infty} \mu_n(U) = \mu(U)$.
d) For any Borel set $B \subset \mathbb{R}$ such that $\mu(\partial B) = 0$, $\lim_{n\to\infty} \mu_n(B) = \mu(B)$ (recall that ∂B is the boundary of B).

Note that condition d) of the theorem can in turn be rephrased into the following simpler one: (μ_n) converges weakly to μ if and only if for all $a < b$ continuity points of the distribution function F_μ,

$$\lim_{n\to\infty} \mu_n(\,]a, b[\,) = \mu(\,]a, b[\,). \tag{1.5.1}$$

Weak convergence allows us to define a further notion of convergence for sequences of random variables.

Definition 1.5.4. *A sequence of random variables (X_n) (not necessarily defined on the same probability space) is said to converge in distribution (or in law) to a random variable X (and this is denoted by $X_n \xrightarrow{d} X$) if the sequence of distributions (μ_{X_n}) converges weakly to μ_X.*

Let us note that convergence in probability implies convergence in distribution. Reciprocally, we have the following proposition.

Proposition 1.5.5. *Let (μ_n) be a sequence of distributions converging weakly to μ. Then there exists a sequence of random variables (X_n) and X defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mu_n = \mu_{X_n}$, $\mu = \mu_X$ and $X_n \to X$ a.s.*

1.5.1 Weak convergence and distribution function

One deduces easily from (1.5.1) that the sequence (μ_n) converges weakly to μ, according to definition 1.5.1, if and only if

$$\lim_{n\to\infty} F_{\mu_n}(t) = F_\mu(t), \quad \text{for all } t \text{ continuity points of } F_\mu.$$

This condition is actually the easiest to check in many practical situations. We shall see other nice characterizations of weak convergence in the following paragraphs.

A slightly weaker notion of convergence for distribution functions is that of vague convergence, which we define below.

Definition 1.5.6. *A sequence (F_n) of distribution functions is said to converge vaguely to a non-decreasing and right-continuous function F on $\mathbb{R}$ if*

$$\lim_{n\to\infty} F_n(t) = F(t), \quad \textit{for all } t \textit{ continuity points of } F.$$

The only difference with the preceding definition is that F may not be a distribution function. The reason for being interested in vague convergence comes from the next theorem, known as Helly's selection theorem.

Theorem 1.5.7. *(Helly)*
Any sequence (F_n) of distribution functions admits a subsequence $(F_{n(k)})$ that converges vaguely.

When does vague convergence imply weak convergence? In order to answer this question, we need the following notion.

Definition 1.5.8. *A sequence (μ_n) of distributions (resp. a sequence (F_n) of distribution functions) is said to be tight, if for all $\varepsilon > 0$, there exists $M > 0$ such that*

$$\limsup_{n\to\infty} \mu_n([-M,M]^c) < \varepsilon \quad \left(\textit{resp. } \limsup_{n\to\infty} \left(1 - F_n(M) + F_n(-M)\right) < \varepsilon \right),$$

that is, up to a factor ε, all the weight of the distribution μ_n remains in the bounded interval $[-M, M]$ as n goes to infinity.

Proposition 1.5.9. *If (F_n) converges vaguely to F and (F_n) is tight, then F is a distribution function.*

Remark 1.5.10. Recalling the following criterion for the convergence of real numbers:

$$x_n \to x \quad \text{iff any subsequence } (x_{n(k)}) \text{ admits itself a subsequence that converges to } x,$$

we see that

$$\mu_n \Rightarrow \mu \quad \text{iff any subsequence } (\mu_{n(k)}) \text{ admits itself a subsequence that converges weakly to } \mu. \tag{1.5.2}$$

We could therefore be tempted to say, using Helly's theorem 1.5.7, that any tight sequence (μ_n) of distributions converges weakly to some limiting distribution μ. This is wrong however, since nothing guarantees that all subsequences of (μ_n) converge weakly to the same limit.

Example 1.5.11. Let us consider the following sequences of distributions: $\mu_n = \mathcal{U}([-n, n])$ or $\mu_n = \mathcal{N}(0, n)$. It is easy to see in both cases that the corresponding sequences of distribution functions F_{μ_n} satisfies

$$\lim_{n\to\infty} F_{\mu_n}(t) = \frac{1}{2}, \quad \forall t \in \mathbb{R}$$

However, the constant function $F(t) \equiv \frac{1}{2}$ is not a distribution function. As an immediate corollary, we obtain that the sequence (μ_n) is not tight. Actually, one can check the stronger statement that for any $M > 0$,

$$\limsup_{n\to\infty} \mu_n([-M, M]^c) = 1,$$

that is, all the weight of the limiting distribution is "lost at infinity".

1.5.2 Weak convergence and characteristic function

The following theorem is known as Lévy's continuity theorem.

Theorem 1.5.12. *(Lévy)*
Let (μ_n) be a sequence of distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

a) If (μ_n) converges weakly to a limiting distribution μ, then the corresponding sequence (ϕ_{μ_n}) of characteristic functions converges pointwise to ϕ_μ.

b) Reciprocally, if the sequence (ϕ_{μ_n}) converges pointwise to a limiting function ϕ that is continuous at 0, then the sequence (μ_n) is tight, ϕ is a characteristic function and (μ_n) converges weakly to the distribution μ with characteristic function ϕ.

Proof. (sketch)
Part a) of the theorem is a direct consequence of the definition 1.5.1 of weak convergence, since $x \mapsto e^{itx}$ is a bounded continuous function on $\mathbb{R}$, for any $t \in \mathbb{R}$.

In order to prove part b), we would first have to check that the assumption of ϕ_{μ_n} converging to ϕ continuous at 0 ensures that the sequence (μ_n) is tight (but let us skip this). Therefore, any subsequence $(\mu_{n(k)})$ admits itself a subsequence converging weakly to a distribution ν, and the part a) proved above implies that $\phi_\nu = \phi$, so ϕ is a characteristic function. Fourier's inversion's theorem 1.4.6 then implies that ν is independent of the subsequence considered, and (1.5.2) allows us to conclude. □

One can see why continuity at 0 is needed, by considering the following example. If $\mu_n = \mathcal{N}(0, n)$, then $\phi_{\mu_n}(t) = \exp(-\frac{nt^2}{2})$ and this sequence of characteristic functions converges pointwise to ϕ given by $\phi(0) = 1$ and $\phi(t) = 0$ for all $t \neq 0$. This function is neither continuous at 0 nor a characteristic function.

1.5.3 Central limit theorem

Let (X_n) be a sequence of i.i.d. random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad n \geq 1.$$

We recall below the well known important consequence of Lévy's theorem.

Theorem 1.5.13. *(Central limit theorem)*
If $\mathbb{E}(X_1^2) < \infty$, *then*

$$\sqrt{n}\,(S_n - m) \xrightarrow{d} Y \sim \mathcal{N}(0, \sigma^2).$$

where $m = \mathbb{E}(X_1)$ *and* $\sigma^2 = Var(X_1)$.

This is to say that the sequence of distributions μ_n of $\sqrt{n}\,(S_n - m)$ converges weakly to a gaussian distribution $\mu = \mathcal{N}(0, \sigma^2)$. This result is therefore a refinement of the law of large numbers; it tells us that the deviation of S_n from m is of order $\frac{1}{\sqrt{n}}$ and that the law of this deviation tends to a Gaussian as n goes to infinity. Note that the Gaussian law appears independently of the law of X_1; the central limit theorem is therefore a *universal* result.

Proof. (sketch)
Let us consider the sequence of characteristic functions

$$\begin{aligned}\phi_{\mu_n}(t) &= \mathbb{E}(\exp(it\sqrt{n}\,(S_n - m))) = \mathbb{E}\left(\exp\left(\frac{it}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m)\right)\right) \\ &= \prod_{i=1}^{n}\mathbb{E}\left(\exp\left(\frac{it}{\sqrt{n}}\,(X_i - m)\right)\right) = \left(\phi_1\left(\frac{t}{\sqrt{n}}\right)\right)^n\end{aligned}$$

where ϕ_1 is the characteristic function of the law of $X_1 - m$ (note that we have used here proposition 1.4.7). Since X_1 is square-integrable, we have the following Taylor's expansion ((1.4.3) for $k = 2$):

$$\phi_1(t) = 1 - \frac{\sigma^2\,t^2}{2} + o(t^2).$$

A small piece of analysis (watch out that the above $o(t^2)$ is complex) allows us to conclude that

$$\lim_{n\to\infty}\phi_{\mu_n}(t) = \lim_{n\to\infty}\left(1 - \frac{\sigma^2\,t^2}{2n} + o\Big(\frac{t^2}{n}\Big)\right)^n = \exp\left(-\frac{\sigma^2\,t^2}{2}\right).$$

This function is continuous at 0 and it is the characteristic function of $\mathcal{N}(0, \sigma^2)$, so conclusion follows by Lévy's theorem. □

Homework 1.5.14. (De Moivre-Laplace's theorem)
Prove the central limit theorem in the case where (X_n) is a Bernoulli sequence with $\mathbb{P}(X_1 = +1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$, without the help of Lévy's theorem.

1.5.4 Weak convergence and moments

Theorem 1.5.15. *Let* $(m_k, k \geq 0)$ *be a sequence of real numbers satisfying condition (1.4.4). If* (μ_n) *is a sequence of distributions such that for all* $k \geq 0$,

$$\int_{\mathbb{R}} x^k\,d\mu_n(x) \to m_k, \quad \text{as } n \to \infty,$$

then there exists a unique distribution μ *with sequence of moments* (m_k) *and such that* (μ_n) *converges weakly to* μ.

Proof. The sequence (μ_n) is tight because for any $M > 0$,

$$\mu_n([-M, M]^c) = \int_{[-M,M]^c} 1\,d\mu_n(x) \leq \int_{[-M,M]^c}\frac{x^2}{M^2}\,d\mu_n(x) \leq \frac{1}{M^2}\int_{\mathbb{R}} x^2\,d\mu_n(x)$$

and $\limsup_{n\to\infty} \int_{\mathbb{R}} x^2 \, d\mu_n(x) = m_2 < \infty$ by assumption. Therefore, any subsequence of the sequence (μ_n) admits itself a subsequence that converges weakly to a distribution with the sequence of moments (m_k). Condition (1.4.4) and theorem 1.4.10 then imply that there is only one such distribution, so (1.5.2) allows us to conclude. □

The advantage of this characterization of weak convergence is that only a countable number of limits needs to be computed. On the other hand, it has two serious drawbacks: first, all the moments of all distributions μ_n need to exist, as those of the limiting distribution μ; then, even in the case where we know that the limit exists, it is not always a trivial problem to deduce the distribution μ from the sequence (m_k), even though formulas like (1.4.8) or (1.4.11) may help (used together with inversion's theorems).

1.5.5 Weak convergence and Stieltjes' transform

Theorem 1.5.16. *A sequence (μ_n) of distributions converges weakly to a limiting distribution μ if and only if the corresponding sequence (g_{μ_n}) of Stieltjes' transforms converges to g_μ uniformly on compact sets of $\mathbb{C}_+$.*

Proof. (sketch)
We first prove that weak convergence of (μ_n) implies uniform convergence of (g_{μ_n}) on compact sets of $\mathbb{C}_+$. It is actually sufficient to verify that

$$\sup_{z\in R} |g_{\mu_n}(z) - g_\mu(z)| \to 0, \quad \text{as } n \to \infty. \tag{1.5.3}$$

for all rectangles $R = [-M, M] \times [m, M]$ with $M > m > 0$. Using the Arzela-Ascoli theorem, the above uniform convergence takes place if

$$g_{\mu_n}(z) \to g_\mu(z), \quad \forall z \in R,$$

and for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sup_{n\geq 1} |g_{\mu_n}(z_1) - g_{\mu_n}(z_2)| \leq \varepsilon, \quad \text{whenever } |z_1 - z_2| \leq \delta, \ z_1, z_2 \in R.$$

The first condition is verified since $x \mapsto \frac{1}{x-z}$ is a bounded continuous function on $\mathbb{R}$, for any $z \in R$. Let us check the second:

$$\begin{aligned} |g_{\mu_n}(z_1) - g_{\mu_n}(z_2)| &\leq \int_{\mathbb{R}} \left| \frac{1}{x - z_1} - \frac{1}{x - z_2} \right| d\mu_n(x) = \int_{\mathbb{R}} \left| \frac{z_2 - z_1}{(x - z_1)(x - z_2)} \right| d\mu_n(x) \\ &\leq \frac{|z_2 - z_1|}{m^2} \int_{\mathbb{R}} d\mu_n(x) = \frac{|z_2 - z_1|}{m^2}, \end{aligned}$$

so the conclusion follows, since m is a fixed positive number.

We then give an idea of the proof of the reverse implication, which follows the lines of the proof of Lévy's theorem. The uniform convergence of (g_{μ_n}) on compact sets of $\mathbb{C}_+$ guarantees that (μ_n) is tight (again, this needs to be checked carefully). So every subsequence $(\mu_{n(k)})$ admits itself a subsequence that converges weakly to a distribution ν, and the first part of the proof implies that $g_\nu = g_\mu$. Inversion's theorem 1.4.17 then shows that $\nu = \mu$, independently of the subsequence considered and (1.5.2) allows us to conclude. □

1.6 Empirical distribution and convergence

Definition 1.6.1. *The empirical distribution μ_n of a set of n random variables $X_1, ..., X_n$ is given by*

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

This is to say that

$$\mu_n(A) = \frac{1}{n} \#\{i \in \{1, ..., n\} : X_i \in A\}, \quad \forall A \in \mathcal{B}(\mathbb{R}),$$

where $\#B$ denotes the cardinality of B.

Note that μ_n is a *random* distribution in general and that for any Borel-measurable function $f : \mathbb{R} \to \mathbb{C}$, we have

$$\int_{\mathbb{R}} f(x)\, d\mu_n(x) = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The empirical distribution of a set of i.i.d. random variables converges to the theoretical distribution, as shown in the following proposition.

Proposition 1.6.2. *If (X_n) is a sequence of i.i.d. random variables with law μ, then the sequence of corresponding empirical distributions μ_n converges weakly to μ, almost surely.*

Proof. (sketch)
Using the convergence criterion established for distribution functions, we see that what needs to be proved is that

$$\lim_{n \to \infty} F_{\mu_n}(t) = F_\mu(t) \quad \text{for all } t \text{ continuity points of } F_\mu, \qquad \text{a.s.}$$

There is a technical issue concerning the position of the two letters "a.s." that we do not address here. We simply note that

$$F_{\mu_n}(t) = \frac{1}{n} \#\{i \in \{1, ..., n\} : X_i \leq t\} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq t\}}$$

and that the random variables $Y_i = 1_{\{X_i \leq t\}}$ are i.i.d. with mean

$$\mathbb{E}(Y_1) = \mathbb{P}(X_1 \leq t) = F_\mu(t),$$

so we obtain by the law of large numbers 1.3.9 that

$$\lim_{n \to \infty} F_{\mu_n}(t) = F_\mu(t), \text{ a.s.}$$

for all $t \in \mathbb{R}$. Modulo the above technical issue, the proposition is proved. □

This result is an illustration of the fact that a sequence of random distributions may converge almost surely to a deterministic limit. We will encounter the same phenomenon when considering the empirical distribution of eigenvalues of random matrices.

1.7 Concentration

For large n-dimensional systems, asymptotic results of the type "law of large numbers" often present much interest on their own, even for relatively small values of n. In many applications however, it may happen that one becomes interested in studying the deviation of some quantity depending on n from its asymptotic value. This can be done essentially in two ways, either by studying the law of the standard deviation using a central limit theorem, or by looking at probabilities of large deviations. We review here this second method.

1.7.1 Large deviations principle

Let $(\mu_n,\, n \geq 1)$ be a sequence of distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 1.7.1. *Let $(a_n,\, n \geq 1)$ be an increasing sequence of positive integers (so $a_n \geq n$) and $I : \mathbb{R} \to [0, \infty]$ be a function such that*

$$\{x \in \mathbb{R} : I(x) \leq \alpha\} \text{ is a compact set}, \quad \forall \alpha \in \mathbb{R}_+. \tag{1.7.1}$$

The sequence (μ_n) is said to satisfy a large deviations principle with speed a_n and (good) rate function I if

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(G) \leq - \inf_{x\in G} I(x), \quad \text{for any closed set } G \subset \mathbb{R}, \tag{1.7.2}$$

and

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(U) \geq - \inf_{x\in U} I(x), \quad \text{for any open set } U \subset \mathbb{R}. \tag{1.7.3}$$

In a more readable way, conditions (1.7.2) and (1.7.3) may be rewritten together as

$$\lim_{n\to\infty} \frac{1}{a_n} \log \mu_n(B) = - \inf_{x\in B} I(x), \quad \text{for any "nice" set } B,$$

or even more intuitively as

$$\mu_n(B) \sim \exp\left(-a_n \inf_{x\in B} I(x)\right), \quad \text{as } n \to \infty,$$

which says in particular that if $\inf_{x\in B} I(x) > 0$, then $\mu_n(B)$ decays exponentially to zero as n goes to infinity. Note that the above principle is sharp, because it gives an upper *and* a lower bound to $\mu_n(B)$.

Remark 1.7.2. Condition (1.7.1) implies that the infimum of I on any closed set G is achieved. Sometimes, condition (1.7.1) is replaced by the weaker condition that I is lower semi-continuous, that is,

$$\{x \in \mathbb{R} : I(x) \leq \alpha\} \text{ is a closed set}, \quad \forall \alpha \in \mathbb{R}_+,$$

in which case the rate function I is not anymore "good", because its infimum on a closed set may not be achieved.

In the following, we will essentially be interested in obtaining upper bounds of the type (1.7.2) for various sequences of distributions (μ_n). The upper bound implies that there exists $C > 0$ and n_0 sufficiently large such that

$$\mu_n(G) \leq C \exp\left(-a_n \inf_{x\in G} I(x)\right), \quad \forall n \geq n_0.$$

Consider now the particular case where $I : \mathbb{R} \to [0, \infty]$ is a function with a unique root $x_0 \in \mathbb{R}$. For any $\varepsilon > 0$, the set $]x_0 - \varepsilon, x_0 + \varepsilon[^c$ is closed, so the assumptions made on I and remark 1.7.2 imply that

$$\inf_{x\in\,]x_0-\varepsilon, x_0+\varepsilon[^c} I(x) = I_\varepsilon > 0.$$

Using the upper bound (1.7.2), we therefore obtain that there exists $C > 0$ and n_0 sufficiently large such that

$$\mu_n\left(\,]x_0 - \varepsilon, x_0 + \varepsilon[^c\right) \leq C \exp(-a_n\, I_\varepsilon), \quad \forall n \geq n_0.$$

In other words, all the weight of the distribution μ_n concentrates around x_0 at exponential speed a_n, as n goes to infinity. This concentration phenomenon is of most interest for us.

Homework 1.7.3. Prove the statement made in remark 1.7.2 and find an example of "bad" rate function I for which the statement is wrong.

1.7.2 An example: sequences of i.i.d. random variables

Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let ν be the law of X_1 (and therefore of any X_n). We assume that

$$\psi_\nu(t) = \mathbb{E}(e^{tX_1}) = \int_{\mathbb{R}} e^{tx}\, d\nu(x) < \infty, \quad \text{for some } t > 0 \text{ and some } t < 0, \tag{1.7.4}$$

(note that this condition implies that all moments of ν exist) and define the function $I_\nu : \mathbb{R} \to [0, \infty]$ by

$$I_\nu(x) = \sup_{t \in \mathbb{R}} (xt - \log \psi_\nu(t)), \quad x \in \mathbb{R}.$$

We state the following lemma without proof.

Lemma 1.7.4. *Under assumption (1.7.4), the function I_ν defined above is convex, satisfies (1.7.1) and has a unique root $x_0 = \mathbb{E}(X_1) = \int_{\mathbb{R}} x\, d\nu(x)$. Moreover, it is non-increasing on $]-\infty, x_0]$, non-decreasing on $[x_0, \infty[$ and*

$$I_\nu(x) = \sup_{t \geq 0} (xt - \log \psi_\nu(t)), \quad \forall x \geq x_0, \quad \textit{and} \quad I_\nu(x) = \sup_{t \leq 0} (xt - \log \psi_\nu(t)), \quad \forall x \leq x_0.$$

Let $S_n = \frac{1}{n} \sum_{i=1}^n X_i$ and μ_n be the law of S_n. We have the following theorem, due to Cramér.

Theorem 1.7.5. *(Cramér)*
Under assumption (1.7.4), the sequence of distributions (μ_n) satisfies a large deviations principle with speed n and rate function I_ν. In particular, for all $\varepsilon > 0$, there exists $C > 0$ and n_0 sufficiently large such that

$$\mathbb{P}(|S_n - x_0| \geq \varepsilon) = \mu_n(\,]x_0 - \varepsilon, x_0 + \varepsilon[^c) \leq C \exp\left(-n \inf_{x \in\,]x_0-\varepsilon, x_0+\varepsilon[^c} I_\nu(x)\right), \quad \forall n \geq n_0. \tag{1.7.5}$$

Proof. We shall not prove here the whole large deviations principle; we only prove inequality (1.7.5). First note that

$$\mathbb{P}(|S_n - x_0| \geq \varepsilon) = \mathbb{P}(S_n \geq x_0 + \varepsilon) + \mathbb{P}(S_n \leq x_0 - \varepsilon). \tag{1.7.6}$$

For the first term on the right-hand side, we obtain by Chebychev's inequality that for all $t \geq 0$,

$$\begin{aligned} \mathbb{P}(S_n \geq x_0 + \varepsilon) &\leq e^{-t(x_0+\varepsilon)}\, \mathbb{E}\left(e^{tS_n}\right) = e^{-t(x_0+\varepsilon)}\, \mathbb{E}\left(e^{\frac{t}{n}(X_1+\ldots+X_n)}\right) \\ &= e^{-t(x_0+\varepsilon)} \left(\psi_\nu\left(\frac{t}{n}\right)\right)^n = \exp\left(-n\left(\frac{t}{n}(x_0+\varepsilon) - \log \psi_\nu\left(\frac{t}{n}\right)\right)\right). \end{aligned}$$

Since this inequality is satisfied for all $t \geq 0$, we also have

$$\begin{aligned} \mathbb{P}(S_n \geq x_0 + \varepsilon) &\leq \inf_{t \geq 0} \exp\left(-n\left(\frac{t}{n}(x_0+\varepsilon) - \log \psi_\nu\left(\frac{t}{n}\right)\right)\right) \\ &= \exp\left(-n \sup_{t \geq 0}\left(\frac{t}{n}(x_0+\varepsilon) - \log \psi_\nu\left(\frac{t}{n}\right)\right)\right) \\ &= \exp\left(-n \sup_{t \geq 0} \left(t\,(x_0+\varepsilon) - \log \psi_\nu(t)\right)\right), \\ &= \exp\left(-n\, I_\nu(x_0+\varepsilon)\right), \end{aligned}$$

by lemma (1.7.4). A similar computation shows that

$$\mathbb{P}(S_n \leq x_0 - \varepsilon) \leq \exp\left(-n\, I_\nu(x_0-\varepsilon)\right),$$

so using (1.7.6, we obtain that

$$\mathbb{P}(|S_n - x_0| \geq \varepsilon) \leq 2 \exp\left(-n \inf_{x \in]x_0-\varepsilon, x_0+\varepsilon[^c} I_\nu(x)\right),$$

which concludes the proof. □

This concentration phenomenon has the following immediate consequence:

$$\sum_{n\geq 1} \mathbb{P}(|S_n - x_0| \geq \varepsilon) = \sum_{n\geq 1} \mu_n\left(\,]x_0 - \varepsilon, x_0 + \varepsilon[^c\right) < \infty, \quad \forall \varepsilon > 0,$$

(because of the exponential decrease). Part a) of Borel-Cantelli's lemma 1.3.4 therefore implies that $S_n \to x_0$ a.s. This tells us that we may easily recover "law of large numbers" results from concentration results.

Example 1.7.6. In the case where $\nu = \mathcal{N}(x_0, \sigma^2)$, we have

$$\psi_\nu(t) = \exp\left(\frac{\sigma^2 t^2}{2} + x_0 t\right) < \infty, \quad \forall t \in \mathbb{R},$$

and

$$I_\nu(x) = \frac{(x - x_0)^2}{2\sigma^2}, \quad x \in \mathbb{R}.$$

I_ν is therefore continuous (implying (1.7.1)), convex and has a unique root $x_0 \in \mathbb{R}$. By the proof of the preceding theorem, we obtain that for any $\varepsilon > 0$,

$$\mathbb{P}(|S_n - x_0| \geq \varepsilon) \leq 2 \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right). \tag{1.7.7}$$

Homework 1.7.7. - Prove the assertions made in lemma 1.7.4.
- Check the computations of example 1.7.6.
- Compute ψ_ν and I_ν when ν is the binomial distribution given by $\nu(\{+1\}) = \nu(\{-1\}) = \frac{1}{2}$.

1.7.3 Talagrand's concentration inequalities

What we deduce from the preceding paragraph is that the empirical mean of n i.i.d. random variables concentrates around the theoretical mean with exponential speed n. Talagrand has shown that this concentration phenomenon holds in much greater generality for functions depending on n independent random variables.

Let us recall here that the median of a random variable X is the real number $M(X)$ defined as

$$M(X) = \sup\{t \in \mathbb{R} : \mathbb{P}(X \leq t) \leq \frac{1}{2}\}$$

and that a function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz with Lipschitz constant L if

$$|f(x) - f(y)| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

where $\|x\|^2 = x_1^2 + ... + x_n^2$.

Theorem 1.7.8. *(Talagrand)*
If $(X_1, .., X_n)$ is a collection of independent random variables such that $\max_{1\leq i\leq n} |X_i| \leq K$ a.s. and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and Lipschitz with Lipshitz constant L, then for all $\varepsilon > 0$,

$$\mathbb{P}(|f(X_1, ..., X_n) - M(f(X_1, ..., X_n))| \geq \varepsilon) \leq 4 \exp\left(-\left(\frac{\varepsilon}{4KL}\right)^2\right).$$

Although it is not directly clear from the above formulation, an important feature of this theorem is that the resulting speed of concentration is equal to the number n of independent random variables involved in the problem. We shall come back to this point when dealing with eigenvalues of random matrices, but let us illustrate this in a well known situation.

Let us consider for instance the case where the X_i are i.i.d., bounded by some constant K and

$$f(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

so that $f(X_1, ..., X_n) = S_n$, the empirical mean of the X_i. f is convex because it is linear and

$$|f(x) - f(y)| \leq \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \leq \frac{1}{n} \sqrt{\sum_{i=1}^{n} 1} \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} = \frac{1}{\sqrt{n}} \|x - y\|,$$

so f has Lipschitz constant $L = \frac{1}{\sqrt{n}}$. The above theorem implies therefore that

$$\mathbb{P}(|S_n - M(S_n)| \geq \varepsilon) \leq 4 \exp\left(-\frac{n\varepsilon^2}{16K^2}\right),$$

which is of the same flavor as inequality (1.7.7). There are slight differences however:

- The expectation $\mathbb{E}(X_1)$ has been replaced by the median $M(S_n)$. Note that either the above result itself or the law of large numbers (see section 1.3.1) imply that $M(S_n)$ tends to $\mathbb{E}(X_1)$ as n goes to infinity. This remains true for many situations where concentration takes place.

- Although the above result is valid only for bounded random variables, it can be extended to more general random variables such as gaussian random variables (and the assumption that f is convex can also be removed).

Bibliography

[1] Bai, Z. D. Methodologies in spectral analysis of large-dimensional random matrices, a review. Statist. Sinica 9 (1999), no. 3, 611–677.

[2] Bai, Z. D. Circular law. Ann. Probab. 25 (1997), no. 1, 494–529.

[3] Baik, J.; Deift, P.; Johansson, K. On the distribution of the length of the longest increasing subsequence of random permutations. J. Amer. Math. Soc. 12 (1999), no. 4, 1119–1178.

[4] Ben Arous, G.; Guionnet, A. Large deviations for Wigner's law and Voiculescu's non-commutative entropy. Probab. Theory Related Fields 108 (1997), no. 4, 517–542.

[5] Billingsley, P. Probability and measure. Third edition. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1995.

[6] Deift, P. A. Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. Courant Lecture Notes in Mathematics, 3. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999.

[7] Dembo, A.; Zeitouni, O. Large deviations techniques and applications. Second edition. Applications of Mathematics (New York), 38. Springer-Verlag, New York, 1998.

[8] Durrett, R. Probability: theory and examples. Second edition. Duxbury Press, Belmont, CA, 1996.

[9] Dyson, F. J. Statistical theory of the energy levels of complex systems. J. Mathematical Phys. 3 (1962), 140–175.

[10] Dyson, F. J. Correlations between eigenvalues of a random matrix. Comm. Math. Phys. 19 (1970), 235–250.

[11] Edelman A. Eigenvalues and Condition Numbers of Random Matrices. MIT PhD Dissertation, 1989. http://math.mit.edu/∼edelman/thesis/thesis.ps

[12] Edelman, A. Eigenvalues and condition numbers of random matrices. SIAM J. Matrix Anal. Appl. 9 (1988), no. 4, 543–560.

[13] Edelman, A.; Kostlan, E.; Shub, M. How many eigenvalues of a random matrix are real? J. Amer. Math. Soc. 7 (1994), no. 1, 247–267.

[14] Emery, M.; Nemirovski, A.; Voiculescu, D. Lectures on probability theory and statistics. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998. Edited by Pierre Bernard. Lecture Notes in Mathematics, 1738. Springer-Verlag, Berlin, 2000.

[15] Fisher, R. A. The sampling distribution of some statistics obtained from non-linear equations. Ann. Eugenics 9 (1939), 238–249.

[16] Ginibre, J. Statistical ensembles of complex, quaternion, and real matrices. J. Mathematical Phys. 6 (1965), 440–449.

[17] Girko, V. L. Theory of random determinants. Mathematics and its Applications (Soviet Series), 45. Kluwer Academic Publishers Group, Dordrecht, 1990.

[18] Girshik M. A. On the sampling theory of the roots of determinantal equations. Ann. Math. Stat. 10 (1939), 203–204.

[19] Goldsheid, I. Y.; Khoruzhenko, B. A. Eigenvalue curves of asymmetric tridiagonal random matrices. Electron. J. Probab. 5 (2000), no. 16, 28 pp. (electronic).

[20] Goodman, N. R. Statistical analysis based on a certain multivariate complex Gaussian distribution. An introduction. Ann. Math. Statist. 34 (1963), 152–177.

[21] Gray R. M. On the asymptotic eigenvalue distribution of Toeplitz matrices. IEEE Trans. Inform. Theory 18 (1972), 725–730.

[22] Grenander, U.; Szegö, G. Toeplitz forms and their applications. Second edition. Chelsea Publishing Co., New York, 1984.

[23] Guionnet, A.; Zeitouni, O. Concentration of the spectral measure for large matrices. Electron. Comm. Probab. 5 (2000), 119–136. http://www.math.washington.edu/∼ejpecp/

[24] Gupta, A. K.; Nagar, D. K. Matrix variate distributions. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, 104. Chapman & Hall/CRC, Boca Raton, FL, 2000.

[25] Haagerup, U. On Voiculescu's R- and S-transforms for free non-commuting random variables. Free probability theory (Waterloo, ON, 1995), 127–148, Fields Inst. Commun., 12, Amer. Math. Soc., Providence, RI, 1997.

[26] Hiai, F.; Petz, D. The semicircle law, free random variables and entropy. Mathematical Surveys and Monographs, 77. American Mathematical Society, Providence, RI, 2000.

[27] Hsu, P. L. On the distribution of roots of certain determinantal equations. Ann. Eugenics 9, (1939), 250–258.

[28] James, A. T. Distributions of matrix variates and latent roots derived from normal samples. Ann. Math. Statist. 35 (1964), 475–501.

[29] Khoruzhenko, B. A.; Khorunzhy, A.; Pastur, L. A.; Shcherbina, M. V. Large-n limit in the statistical mechanics and the spectral theory of disordered systems. In: Domb, C.; Lebowitz J.L. (eds) Phase transitions and critical phenomena, 15. Academic Press Inc., New-York, NY, 1992, 74–239.

[30] Khorunzhy, A. M.; Pastur, L. A. On the eigenvalue distribution of the deformed Wigner ensemble of random matrices. Spectral operator theory and related topics, 97–127, Adv. Soviet Math., 19, Amer. Math. Soc., Providence, RI, 1994.

[31] Kunz, H.; Souillard, B. Sur le spectre des opérateurs aux différences finies alatoires. (French) [On the spectra of random finite difference operators] Comm. Math. Phys. 78 (1980/81), no. 2, 201–246.

[32] Marčenko, V. A.; Pastur, L. A. Distribution of eigenvalues in certain sets of random matrices. Math. USSR-Sbornik 1 (1967), no. 4, 457–483.

[33] Mehta, M. L. Random matrices. Second edition. Academic Press, Inc., Boston, MA, 1991.

[34] Mehta M. L., Gaudin M., On the density of eigenvalues of a random matrix, Nuclear Phys. 18 (1960), 420–427.

[35] Molchanov, S. A.; Pastur, L. A.; Khorunzhy, A. M. Distribution of the eigenvalues of random band matrices in the limit of their infinite order. Theoret. and Math. Phys. 90 (1992), no. 2, 108–118.

[36] Montgomery, H. L. Distribution of the zeros of the Riemann zeta function. Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 1, pp. 379–381. Canad. Math. Congress, Montreal, Que., 1975.

[37] Muirhead, R. J. Aspects of multivariate statistical theory. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1982.

[38] Müller, R. A random matrix model of communication via antenna arrays. IEEE Trans. Inform. Theory 48 (2002), no. 9, 2495–2506.

[39] Odlyzko, A. M. On the distribution of spacings between zeros of the zeta function. Math. Comp. 48 (1987), no. 177, 273–308.

[40] Pastur, L. A. The spectrum of random matrices. (Russian) Teoret. Mat. Fiz. 10 (1972), no. 1, 102–112.

[41] Pastur, L.; Figotin, A. Spectra of random and almost-periodic operators. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 297. Springer-Verlag, Berlin, 1992.

[42] Pastur, L.; Vasilchuk, V. On the law of addition of random matrices. Comm. Math. Phys. 214 (2000), no. 2, 249–286.

[43] Ross, S. M. Initiation aux probabilités. Presses Polytechniques et Universitaires Romandes, Lausanne, 1996.

[44] Roy S.N. p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. Sankhyä 4 (1939), 381–396.

[45] Silverstein, J. W. Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. J. Multivariate Anal. 55 (1995), no. 2, 331–339.

[46] Silverstein, J. W.; Bai, Z. D. On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. J. Multivariate Anal. 54 (1995), no. 2, 175–192.

[47] Speicher, R; Free convolution and the random sum of matrices. Publ. Res. Inst. Math. Sci. 29 (1993), no. 5, 731–744.

[48] Talagrand, M. A new look at independence. Ann. Probab. 24 (1996), no. 1, 1–34.

[49] Telatar, I. E. Capacity of multi-antenna Gaussian channels. European Trans. on Telecom. 10 (1999), no. 6, 585–595.

[50] Telatar, I. E.; Tse, D. N. C. Capacity and mutual information of wideband multipath fading channels. IEEE Trans. Inform. Theory 46 (2000), no. 4, 1384–1400.

[51] Tracy, C. A.; Widom H. Level-spacing distributions and the Airy kernel. Comm. Math. Phys. 159 (1994), no. 1, 151–174.

[52] Tracy, C. A.; Widom, H. Level spacing distributions and the Bessel kernel. Comm. Math. Phys. 161 (1994), no. 2, 289–309.

[53] Tse, D. N. C.; Hanly, S. V. Linear multiuser receivers: effective interference, effective bandwidth and user capacity. IEEE Trans. Inform. Theory 45 (1999), no. 2, 641–657.

[54] Tse D. N. C.; Zeitouni O. Linear multiuser receivers in random environments. IEEE Trans. Inform. Theory 46 (2000), no. 1, 171–188.

[55] Tulino, A. M.; Verdú, S. Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory 1 (1), Now Publishers Inc., 2004.

[56] Verdú, S. Spectral efficiency in the wideband regime. Special issue on Shannon theory: perspective, trends, and applications. IEEE Trans. Inform. Theory 48 (2002), no. 6, 1319–1343.

[57] Voiculescu, D.; A strengthened asymptotic freeness result for random matrices with applications to free entropy. Internat. Math. Res. Notices 1998, no. 1, 41–63.

[58] Voiculescu, D. V.; Dykema, K. J.; Nica, A. Free random variables. A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups. CRM Monograph Series, 1. American Mathematical Society, Providence, RI, 1992.

[59] Wigner E. P. Random matrices in physics, SIAM Review 9 (1967), 1–123.

[60] Wigner E. P. Characteristic vectors of bordered matrices with infinite dimensions I. Ann. Math. (2) 62 (1955), 548–564.

[61] Wigner E. P. Characteristic vectors of bordered matrices with infinite dimensions II. Ann. Math. (2) 65 (1957), 203–207.

[62] Wigner E. P. On the distribution of the roots of certain symmetric matrices. Ann. Math. (2) 67 (1958), 325–327.

[63] Wilf, H. S. Finite sections of some classical inequalities. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 52 Springer-Verlag, New York-Berlin 1970.

[64] Wishart J. The generalised product moment distribution in samples from a normal multivariate population. Biometrika A20 (1928), 32–52.

[65] Zheng, L.; Tse, D. N. C. Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. IEEE Trans. Inform. Theory 48 (2002), no. 2, 359–383.